HOCHSCHULE
ESSLINGEN

Informatik und
Informationstechnik

IT Innovationen

Band 34

Januar 2025

Grußwort der Fakultät

Liebe Leserinnen und Leser,

Als Kind war ich begeisterter Zuschauer der Sendung mit der Maus. Besonders hatten es mir die Erklärgeschichten angetan. Da sah man wie Zahnpasta hergestellt wird oder wie man ein Flugzeug baut… und ich hab‘ mich immer gefragt: wie kann man sowas kompliziertes bauen? Wer denkt sich all diese Dinge aus? Heute weiß ich es. Forscher und Entwickler.

Und auch wenn das Ergebnis bei der Sendung mit der Maus so einfach und selbstverständlich aussieht, das habe ich auch lernen dürfen… einfach und selbstverständlich ist es ganz und gar nicht.

Wenn ich mir heute diese Ausgabe der IT-Innovationen anschaue, dann habe ich ähnliche Gedanken, quasi Sendung-mit-der-Maus-Maus-Vibes. Hier sind Erklärgeschichten für Technologien, die wir bald ganz selbstverständlich nutzen werden. Da ist zum Beispiel die Entwicklung eines Chatbots, der Ihnen beim Autokauf so charmant wie klug zur Seite steht. Oder es wird die Frage beantwortet: Wie kann man einem Fahrzeug beibringen, mit anderen Autos zu “sprechen”, um den Verkehr sicherer zu machen und wie finden diese Autos dann auch effizient einen Parkplatz? Wie macht man aus Handyfotos täuschend echte 3D-Modelle? Klingt wie Science-Fiction? Ist es heute nicht mehr! Es ist das Resultat der harten Arbeit unserer Studierenden und jeder Menge Neugier.

Wir wünschen Ihnen viel Spaß beim Lesen, Staunen und Entdecken. Vielleicht erfahren Sie etwas, das Sie schon immer wissen wollten – oder von dem Sie nicht mal wussten, dass Sie es interessiert. Egal wie: Wir versprechen jede Menge Aha-Momente - leider aber ohne Maus und Elefant.

Viel Freude beim Lesen wünscht Ihnen

Ihr Prof. Dr. Tobias Heer, Dekan

IMPRESSUM

ERSCHEINUNGSORT

73732 Esslingen am Neckar

HERAUSGEBER

Prof. Dr. Tobias Heer
Dekan der Fakultät Informatik und Informationstechnik
der Hochschule Esslingen - University of Applied Sciences

REDAKTIONSANSCHRIFT

Hochschule Esslingen - University of Applied Sciences
Fakultät Informatik und Informationstechnik
Flandernstraße 101
73732 Esslingen am Neckar

Telefon	+49(0)711.397-4210
Telefax	+49(0)711.397-4214
E-Mail	it@hs-esslingen.de
Website	www.hs-esslingen.de/it

REDAKTION, DESIGN, LAYOUT und SATZ

Dipl.-Inform.(FH) Rolf Gassner

Hochschule Esslingen - University of Applied Sciences
Fakultät Informatik und Informationstechnik
Flandernstraße 101
73732 Esslingen am Neckar

ERSCHEINUNGSWEISE

Einmal pro Semester, jeweils Januar und Juni

ISSN 1869-6457

Anbindung einer Sensor Einheit an einen Multicore Prozessor und Verarbeitung der Sensor Daten

Deniz-Can Acici **Michael Scharf**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Leuze electronic GmbH + Co. KG, Owen

Einleitung

Bei vielen Entwicklungsprozessen sollten am Anfang gewisse Anforderungen gesetzt werden. Dabei können sich während des ganzen Prozesses die Anforderungen ändern. Änderungen von solchen Anforderungen kann oftmals eine erhöhte Performance benötigen. Für einen Sensor wurde in der Vergangenheit ein passender Prozessor ausgewählt. Aufgrund geänderter Anforderungen und Preise soll jedoch als Alternative ein neuer Prozessor verwendet werden und in das bestehende Gesamtsystem integriert werden. Der alte Prozessor ist ein Dual-Core Prozessor mit weniger Kernen zur Datenverarbeitung. Der neue Prozessor hat mehr Cores, und sogar mehr Hardware beschleunigende Einheiten zur Echtzeit Verarbeitung von Daten und ist günstiger.

Motivation

Um das Wechseln der Prozessoren zu vereinfachen, wird im Umfang der Arbeit nachgeforscht, wie und ob dieser Umstieg möglich ist. Mithilfe der Hardware beschleunigenden Einheiten soll die Verarbeitung der Sensordaten schneller und parallel zum Hauptprozess laufen. Mit dem neuen Prozessor kann so mehr Platz für weitere Funktionen gewonnen werden und für den Kunden ein schnellerer Sensor verfügbar gestellt werden, ohne ersichtliche Änderung nach außen. So könnte der Sensor ausgetauscht werden, ohne dafür beim Kunden Anpassungen zu benötigen. Dazu ist bei der Entwicklung der neue Prozessor günstiger und bietet mehr Leistung und Funktion.

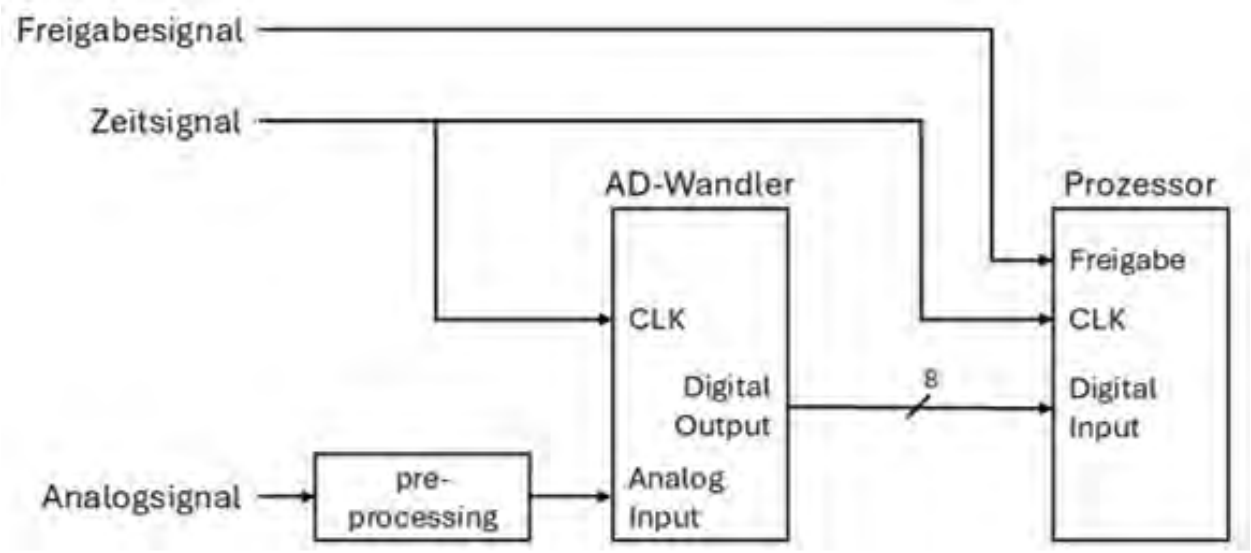


Abb. 1: Vereinfachtes Schaltbild vom Weg der analogen Daten zum Prozessor [1]

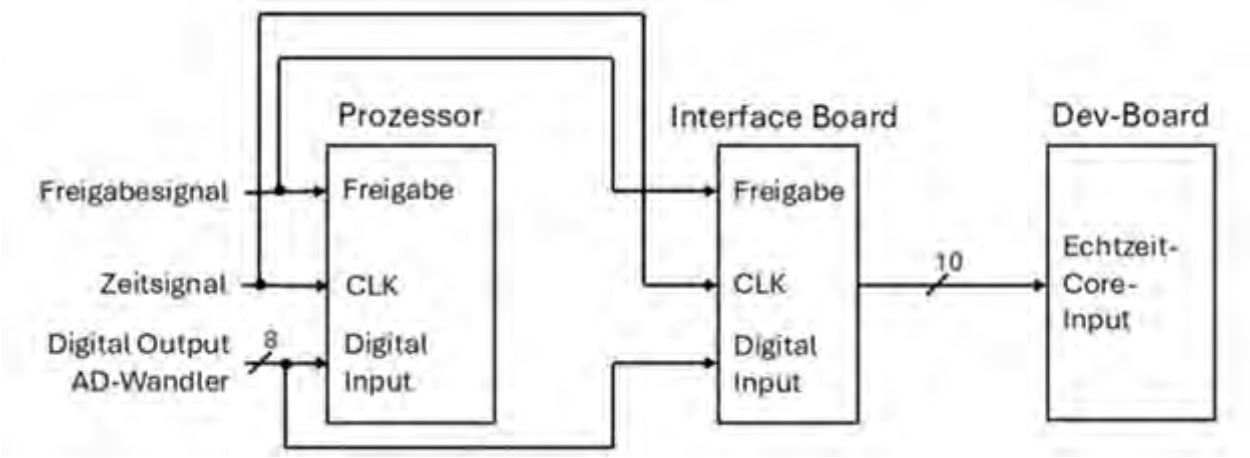


Abb. 2: Prozessor und Developer Board verbunden mit einem Interface Board [1]

Arbeitsbeschreibung

Bei einem Sensor Produkt von Leuze soll ein Chip ausgetauscht werden. Der Chip vom Produkt ist ein Dual-Core Chip mit einer programmierbaren Echtzeiteinheit. Dieser Chip bekommt von einem AD-Wandler momentane Daten. Dazu bekommt der Chip auch noch ein Zeitsignal und ein Freigabesignal. Dieser Chip soll jetzt mit einem neuen Chip ausgetauscht werden. Der neue Chip ist jetzt ein Octa-Core Prozessor mit programmierbaren Echtzeiteinheiten. Dieser Chip soll auf dem Sensor Produkt von Leuze eingebaut werden. Auf der 1 ist abgebildet, wie die analogen Signale zum Prozessor über einen AD-Wandler gehen. Dabei kommen beim Prozessor zwei Signale und ein Datenbündel an. Das Zeitsignal ist ein sich periodisch wiederholendes Signal. Über dieses Signal können verschiedene Prozesse den richtigen zeitlichen Ablauf absichern. Das Freigabesignal gibt an, wann die Analogsignale aufgenommen werden sollen. Die Analogsignale werden vorverarbeitet und dann an den AD-Wandler weitergeleitet, dabei erstellt der AD-Wandler aus einem analogen Signal ein digitales Signal, von 0 bis 255, welches über 8 Bits als Bündel an den Prozessor weitergegeben wird. Damit die Daten schneller verarbeitet werden können, wird der Input von dem Echtzeit-Core genutzt. Dabei kann unabhängig von dem Hauptkernen die ankommenden Daten vorverarbeitet werden. Das Echtzeit-Subsystem wird herkömmlicherweise auf C programmiert, kann aber auch mit Assembler programmiert werden. Wir nutzen C als Programmiersprache. Da C eine hohe Ausführungsgeschwindigkeit hat, wird es oftmals in eingebetteten Systemen genutzt. [2] Auch hier ist dies von Vorteil, soll aber auch genutzt werden da es von den Chipherstellern so vorgeschrieben ist. Damit getestet werden kann, wie der neue Prozessor mit den Daten umgehen kann, wird ein Entwicklungsboard verwendet. Dieses Entwicklungsboard wird mit einem Interfaceboard verbunden, wie auch in 2 zu sehen ist. Dabei werden vom Leuze Produkt die Signale abgezweigt und mit dem Interface Board verbunden, welches dann zum Entwicklungsboard mit einer Schnittstelle verbunden wird. Der Hauptaufwand der Arbeit besteht dann, zu testen, wie die Daten von der Programmierbaren Echtzeiteinheit zum Linux User Space gesendet werden können. Dazu wird ein Mailbox-System verwendet, welches über den Ram die Daten zwischenspeichert. Wenn die Daten im Ram sind, wird dies über dieses Mailbox-System der jeweiligen „Mailbox“ signalisiert, um mitzuteilen, dass die Daten verfügbar sind. Eine vereinfachte Abbildung ist in 3 zu sehen.

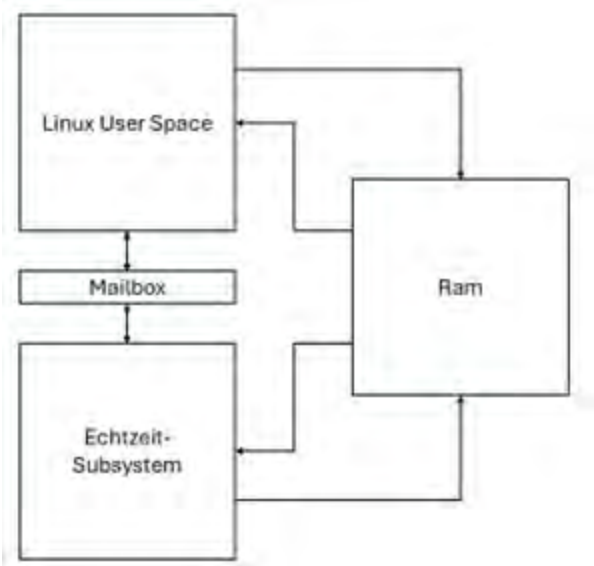


Abb. 3: Vereinfachtes Mailbox System [1]

Ausblick

Für die Verwendung des neuen Prozessors im Produkt steht wenig im Weg. Es muss nur noch der vorhandene Code etwas angepasst werden, um den Sub-Prozess verwirklichen zu können. Zwar ist diese Arbeit nur eine Komponente einer vielschichtigen Gesamtarbeit zur Umstellung zum neuen Prozessor, aber dieser Teil scheint keinen weiteren großen

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Die freie Enzyklopädie Wikipedia. C (Programmiersprache). https://de.wikipedia.org/w/index.php?title=C_(Programmiersprache)&oldid=250069836, 2024.

Entwicklung und Implementierung eines Nutzerkonzepts für Softwaregestützte Topologieoptimierung von Leichtbaustrukturen in Produkt- und Fahrzeugentwicklungsprozessen

Richmore Aidams **Jörg Nitzsche**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Deutsches Zentrum für Luft- und Raumfahrt, Stuttgart

Abstract
Das Institut für Fahrzeugkonzepte des Deutschen Zentrums für Luft- und Raumfahrt (DLR) bearbeitet und koordiniert verkehrsrelevante Forschungsthemen zu neuen Fahrzeugkonzepten und -technologien. Im Projekt Next Generation Train (NGT) werden die Forschungsarbeiten rund um das Schienenfahrzeug gebündelt. Dazu gehören Forschungsthemen im Bereich der Fahrzeugarchitektur für Schienenfahrzeuge. Hier kommen Techniken wie die Topologieoptimierung (TO) zum Einsatz. Das Institut verwendet TO und Finite Elemente (FE) Methoden, um Formen mit algorithmischen Modellen zu optimieren, die das bestmögliche Materiallayout basierend auf benutzerdefinierten Lasten, Bedingungen und Einschränkungen ermitteln.

1. Einleitung

Die Analyse und Weiterverwendung von Optimierungsergebnissen aus der Topologieoptimierung ist in der Regel eine komplexe Herausforderung, die mit Hilfe geeigneter Software gelöst werden kann. Als Softwarelösung wird intern ein Teilautomatisiertes Konstruktions- und Entwicklungstool (TAKT) entwickelt und eingesetzt. TAKT ist eine Python-Anwendung bzw. ein Konstruktionswerkzeug, das speziell auf die Bedürfnisse von Ingenieuren in der Schienenfahrzeugbranche zugeschnitten ist. Die Kernaufgabe darin besteht, aus den Optimierungsergebnissen der Topologieoptimierung unter bestimmten Randbedingungen, wie z.B. der Dichteschwelle, automatisierte Vorschläge für Fachwerkstrukturen mit Balkenelementen zu generieren. Die Aufgaben von TAKT umfassen die Erstellung von FE-, Voxel-, Drahtgitter- und Querschnittsmodellen(vgl. Abbildung 1).

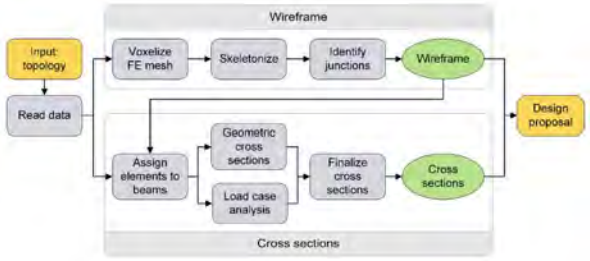


Abb. 1: Automatisierter Prozess zur Ableitung von Designvorschlägen aus Topologieoptimierungen [1]

Anwendungsszenarien von TAKT
In typischen Anwendungsszenarien werden aus den TO-Ergebnissen als Zwischenschritte ein Voxel- und ein Drahtgittermodell und als Endergebnis ein Querschnittsmodell erzeugt. Während des Prozesses können Parameter wie z.B. die Dichteschwelle beim Laden eines neuen Voxelmodells oder die Struktur eines Drahtgittermodelles verändert werden, um Designvorschläge zu modifizieren (vgl. Abbildung 2). Im Allgemeinen kann der Gesamtprozess in zwei Schritten dargestellt werden: Extraktion eines Drahtgittermodells aus der TO und Ableitung eines Querschnittsmodells daraus. Aus beiden Modellen können Designvorschläge abgeleitet werden.

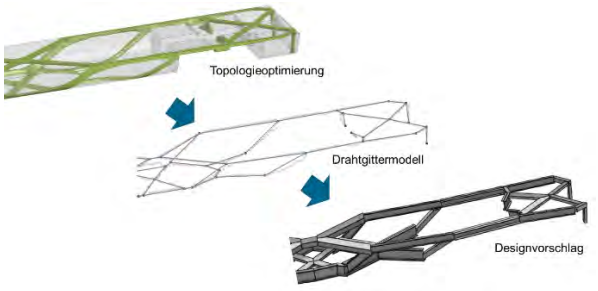


Abb. 2: Vereinfachte Darstellung des TAKT-Verfahrens [1]

2. Problemstellung

Die Problemstellung hier ist, dass die Nutzerführung von TAKT eine Herausforderung für neue Benutzer darstellt. Der Grund dafür ist, dass die effektive Nutzung der Anwendung eine umfassende Einarbeitungszeit erfordert, um die richtigen Konfigurationen und Einstellungen für spezifische Projekte auszuwählen. Dies beinhaltet das Verständnis komplexer Parameter und Optionen, die für die Optimierung der Strukturen entscheidend sind.
Im aktuellen Verfahren erhalten Benutzer Unterstützung durch eine detaillierte Texteinführung, die die grundlegenden Funktionen und Anwendungsbeispiele von TAKT beschreibt. Diese Dokumentation bietet eine wertvolle Orientierung, reicht jedoch möglicherweise nicht aus, um alle Fragen neuer Benutzer zu klären. Daher könnte es erforderlich sein, zusätzliche Schulungsmaterialien oder interaktive Hilfen bereitzustellen, um den Einarbeitungsprozess zu erleichtern und die Benutzerfreundlichkeit der Anwendung zu verbessern.

3. Zielsetzung dieser Arbeit

Diese Arbeit beschäftigt sich mit der Entwicklung und Implementierung eines Nutzerkonzeptes für TAKT. Als Lösung wird das Nutzerkonzept in einer neu gestalteten Benutzeroberfläche umgesetzt. Die Ergebnisse sollen zeigen, dass eine intuitive Navigation und klare Visualisierungen entscheidend für die erfolgreiche Nutzung des Tools sind. Darüber hinaus soll die Arbeit wertvolle Erkenntnisse darüber liefern, wie ein nutzerzentriertes Design den Designprozess beschleunigen und die Fehlerquote in der Fahrzeugentwicklung reduzieren kann.

4. Entwicklung des Konzepts

Das Konzept wird zunächst so entwickelt, dass es bestehende Anwendungsszenarien abdeckt. Eine MosCow-Analyse zeigt, dass die Prioritäten bei der Benutzerfreundlichkeit und der Zentralisierung der Anwendung, z.B. auf einem institutseigenen Server, liegen. Aufgrund dessen besteht das Grundkonzept darin, einführende Schritte, z.B. durch Info-Boxen, in der Benutzeroberfläche zu implementieren, um die Usability der Anwendung zu erhöhen. Zusätzlich werden User-Touchpoints definiert und implementiert. Diese sollen aufzeigen, wann eine Interaktion des Nutzers notwendig ist. Ziel ist es hierbei, Fehler durch Nutzerinteraktionen während des gesamten Prozesses zu minimieren. Zur Umsetzung des Grundkonzeptes werden verschiedene Anwendungsszenarien skizziert und in drei Automatisierungsstufen eingeteilt, wobei in Stufe 1 der Prozess vollständig automatisiert abläuft und in Stufe 3 der Nutzer in den Prozess eingebunden wird. Alle Automatisierungsstufen erzeugen wiederverwendbare Zwischenergebnisse, die abhängig von der Konfiguration einzeln oder gesammelt entweder lokal oder in einer zentralen Datenbank gespeichert werden.

5. Umsetzung des Konzepts

Das Konzept wird mit Hilfe von zwei Frameworks umgesetzt: Vite und FastAPI. Vite wird in Verbindung mit React, einer JavaScript-Bibliothek, verwendet, um das Benutzerkonzept in einer Benutzeroberfläche umzusetzen. Der Vorteil hierbei ist, dass Vite ein leichtgewichtiges Framework ist, welches die Entwicklung und Wartung des Codes vereinfacht. Außerdem können mit React wiederverwendbare Komponenten entwickelt werden. Für das Backend wird FastAPI verwendet, da es ebenfalls ein leichtgewichtiges Framework bietet. Zudem generiert FastAPI automatisch API-Dokumentationen, was aus Entwicklersicht Zeit spart. Abschließend wird Postgres verwendet, um eine Datenbank zur Speicherung der Modelle zu erstellen. Das Ergebnis sind drei Services, die in Docker containerisiert werden.

6. Zusammenfassung

Zusammenfassend kann gesagt werden, dass diese Arbeit die Analyse und Weiterverwendung von Optimierungsergebnissen aus der Topologieoptimierung vereinfacht. Darüber hinaus eröffnet diese Arbeit neue Möglichkeiten, wie z.B. die Durchführung einer Parameterstudie zur Simulation des Einflusses bestimmter Randbedingungen auf neue Designvorschläge.

Literatur und Abbildungen

[1] Christian Gomes Alves, Yannick Barthel, and Matthias Halsner. Automated Derivation of CAD Designs from Topology Optimization Results. *World Congress on Railway Research (WCRR)*, 2022.

Hardware der Intelligenten Parkraumüberwachung

Abdullah Akcay **Clemens Klöck**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

Die intelligente Parkraumerkennung hilft, Parkplätze effizient zu nutzen und die Suche nach freien Stellplätzen zu verkürzen. In urbanen Gebieten führt die begrenzte Anzahl an Parkplätzen oft zu Stress, unnötigem Verkehr und erhöhtem CO2-Ausstoß. Technische Probleme bei Parkscheinautomaten und mangelhafte Zahlungsmöglichkeiten erschweren die Nutzung von Parkraummanagementsystemen (PRM). Auch die ungleiche Auslastung von Parkplätzen und die steigende Nachfrage stellen Betreiber vor Herausforderungen. Intelligente Systeme zur Parkraumerkennung bieten eine Lösung, indem sie die Parkplatzsuche optimieren und die Umweltbelastung reduzieren.

Ziel des Projekts

Dieses Projekt zielt darauf ab, ein Hardware-Modul zu entwickeln, das Parkräume in Echtzeit erkennt und die erfassten Daten an ein übergeordnetes System weiterleitet. Das System soll die Verfügbarkeit von Parkplätzen effizient überwachen und die Daten für Anwendungen wie Parkleitsysteme, Verkehrsmanagement und Umweltanalysen bereitstellen.

Aufbau des Hardware-Moduls

Das Hardware-Modul besteht aus:

1. **Eingabehardware:** Raspberry Pi 5 / Raspberry Pi Zero 2 WH oder NVIDIA Jetson Orin Nano
2. **Kamera:** Raspberry Pi Camera Module V2 oder Intel RealSense D435
3. **Kommunikationsmodule:** Wi-Fi und/oder Ethernet für die Datenübertragung
4. **Energieversorgung:** Step-Down-Konverter und Lithium-Ionen-Batterien für mobilen Einsatz

Die unterschiedlichen Varianten bieten Flexibilität. Der Raspberry Pi Zero 2 WH eignet sich für kleinere Systeme mit geringem Stromverbrauch, während der Raspberry Pi 5 größere Rechenlasten und komplexe Bildverarbeitung bewältigen kann.

Funktionsweise

Das Hardware-Modul erfasst mithilfe der Kamera die Parkplatzsituation. Die Daten werden entweder lokal verarbeitet oder über eine drahtlose Verbindung an ein übergeordnetes System weitergeleitet. Durch die Kombination von Bildverarbeitung und maschinellem Lernen können freie Parkplätze zuverlässig erkannt werden.

Kamera und Bildverarbeitung

Die eingesetzten Kameras, wie die Intel RealSense D435 wie in der Abbildung 1 zu sehen ist, bieten eine hohe Präzision bei der Erfassung von Parkplätzen durch die Integration von Tiefen Sensorik. Diese ermöglicht nicht nur die Erkennung freier Stellplätze, sondern auch die Identifikation von Hindernissen oder unbefugten Parkvorgängen. Mit einer Auflösung von 1280 x 720 Pixeln bei Tiefenaufnahmen gewährleistet die Kamera eine zuverlässige Erkennung selbst bei schwierigen Lichtverhältnissen. Die Bildverarbeitung kann durch maschinelles Lernen optimiert werden, um z. B. Falschmeldungen zu reduzieren. [1]

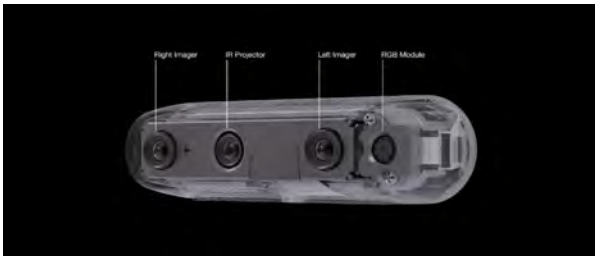


Abb. 1: Intel® RealSense™ D435 Depth Camera [1]

Datenschutz und Sicherheit

Datenschutz spielt eine zentrale Rolle bei der intelligenten Parkraumüberwachung. Um den Anforderungen der Datenschutz-Grundverordnung (DSGVO) gerecht zu werden, erfolgt die Verarbeitung der Bilder ausschließlich lokal auf der Hardware. Es werden keine vollständigen Bilder gespeichert, sondern lediglich anonymisierte Datenmuster oder Tiefeninformationen, die keine Rückschlüsse auf Personen oder Kennzeichen zulassen. Zusätzlich sorgt eine verschlüsselte Datenübertragung über eine SIM Karte für maximale Sicherheit.

Vergleich der Hardware-Optionen

Raspberry Pi Variante
Die Raspberry Pi 5, in der Abbildung 2 ersichtlich, ist eine gute Wahl für unser Parkplatz-Erkennungssystem, da sie die erforderliche Rechenleistung und Flexibilität bietet, um anspruchsvolle Bildverarbeitungsaufgaben effizient zu bewältigen. Ausgestattet mit einem leistungsstarken Quad-Core Cortex-A76 Prozessor und bis zu 8 GB RAM ermöglicht sie die Verarbeitung komplexer Bilddaten in Echtzeit und unterstützt gleichzeitig einfache KI-basierte Berechnungen, die für die präzise Erkennung freier Parkflächen unerlässlich sind. Darüber hinaus bietet der Raspberry Pi 5 eine Vielzahl an Anschlussmöglichkeiten, darunter USB 3.0, Gigabit-Ethernet und zwei HDMI-Ausgänge, die für die Integration in unser System entscheidend sind. Diese Konnektivität erlaubt es, sowohl die Kamera-Module als auch die Übertragungs- und Auswertungskomponenten nahtlos zu verbinden. Im Vergleich zu früheren Modellen liefert der Raspberry Pi 5 die notwendige Leistung, um größere Parkflächen zu überwachen und hochfrequente Datenströme effizient zu verarbeiten. Die Kombination aus Leistungsfähigkeit, Anschlussvielfalt und Skalierbarkeit macht ihn zur besten Wahl für unser Projekt, um die Anforderungen moderner Parkraumerkennungssysteme zu erfüllen. [2]

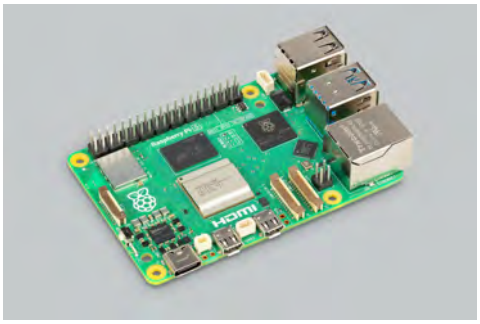


Abb. 2: Der Raspberry Pi 5 - Modell mit grundlegenden Komponenten und Anschlüssen [2]

Der **Raspberry Pi Zero 2 WH**, ebenfalls in der Abbildung 3 dargestellt, ist eine kompakte und kostengünstige Lösung für einfache Anwendungen in der Parkraumerkennung. Mit seinem 64-Bit Quad-Core ARM Cortex-A53 Prozessor und 512 MB RAM bietet er ausreichende Leistung für grundlegende Aufgaben. Dank seiner geringen Größe und des niedrigen Stromverbrauchs ist er ideal für mobile Systeme oder kleinere Parkflächen, bei denen Effizienz und Kosten im Vordergrund stehen. [3]

Abb. 3: Der Raspberry Pi Zero 2 WH - Kompakte Version mit ARM Cortex-A53 Prozessor [3]

NVIDIA Jetson Orin Nano Variante Diese Variante bietet eine leistungsstarke Plattform für Entwickler, die auf künstliche Intelligenz (KI) und maschinelles Lernen angewiesen sind, das Hardware Modul dazu ist in der Abbildung 4 ersichtlich. Es kombiniert eine fortschrittliche 6-Kern-CPU und eine leistungsstarke GPU, die speziell für die effiziente Verarbeitung von KI-Anwendungen und Echtzeit-Bildverarbeitung entwickelt wurde. Dieses Kit eignet sich hervorragend für Anwendungen wie autonome Systeme, Robotik und intelligente Analyse. Der einzige Nachteil ist, dass es zu viel Strom verbrauchen kann und eventuell gut gekühlt werden muss. [4]

Abb. 4: Der NVIDIA Jetson Orin Nano - AI-Entwicklungsplattform [4]

Einsatzmöglichkeiten

- Überwachung von Parkplätzen in Parkhäusern oder offenen Flächen

- Integration in Parkleitsysteme zur Echtzeit-Visualisierung freier Parkplätze
- Verbesserung des Verkehrsflusses in Flughäfen
- Ausgelastete Städte wie New York etc.

Ausblick

Die intelligente Parkraumerkennung bietet großes Potenzial, um Parkflächen effizienter zu nutzen und den Parksuchverkehr zu reduzieren. Zukünftige Entwicklungen könnten die Integration von Sensorfusion, leistungsfähigerer KI und die Nutzung in autonomen Fahrzeugen umfassen, wodurch Echtzeit-Daten für Navigation und Verkehrsmanagement bereitgestellt werden.
Fortschritte in der Bildverarbeitung und neue Sensoren werden die Präzision des Systems weiter steigern. Gleichzeitig bleibt der Datenschutz ein zentraler Aspekt, mit Lösungen wie lokaler Datenverarbeitung und anonymisierten Daten, die modernen Standards gerecht werden.
Dieses System verbindet technologische Vielseitigkeit mit einer nachhaltigen Vision und legt den Grundstein für effizientere, umweltfreundlichere Mobilitätslösungen.

Literatur und Abbildungen

[1] Intel Corporation. Intel RealSense Depth Camera D435. https://www.intelrealsense.com/depth-camera-d435/, 10 2024.

[2] Raspberry Pi Foundation. Raspberry Pi 5 Product Brief. https://datasheets.raspberrypi.com/rpi5/raspberry-pi-5-product-brief.pdf, 08 2024.

[3] Raspberry Pi Foundation. Raspberry Pi Zero 2 W Product Brief. https://datasheets.raspberrypi.com/rpizero2/raspberry-pi-zero-2-w-product-brief.pdf, 04 2024.

[4] Seeed Studio. Jetson Orin Nano Developer Kit Datasheet. https://files.seeedstudio.com/wiki/Jetson-Orin-Nano-DevKit/jetson-orin-nano-developer-kit-datasheet.pdf, 02 2023.

Optimierung eines Large Language Models für einen Chatbot zur personalisierten Fahrzeugkaufberatung

Cihan-Osman Akgoez **Steffen Schober**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma adesso SE, Stuttgart

Einführung

Die steigende Nachfrage nach personalisierten und effizienten Beratungslösungen im Fahrzeugkauf stellt eine Herausforderung dar, die durch moderne Technologien wie Large Language Models (LLMs) bewältigt werden kann. Ziel ist die Entwicklung eines Chatbots, der mithilfe eines LLMs sowohl technische Anforderungen als auch individuelle Nutzerpräferenzen berücksichtigt. LLMs finden vielseitige Anwendungen in der Automobilbranche, beispielsweise in Advanced Driver Assistance Systems (ADAS) [1]. Sie ermöglichen zudem die Analyse großer Datenmengen, einschließlich technischer Spezifikationen und Markttrends, um gezielte Empfehlungen zu generieren Diese Arbeit nutzt Fine-Tuning auf domänenspezifische Datensätze, Retrieval-Augmented Generation [7] (RAG) für aktuelle Informationen und Learning from Human Feedback (RLHF) zur iterativen Optimierung. Die Evaluierung erfolgt durch automatische Metriken wie BLEU und ROUGE [5] sowie durch direktes Nutzerfeedback.

Problemstellung

Die Entwicklung eines Fahrzeugkaufberatungs-Chatbots stellt technische und konzeptionelle Herausforderungen dar. Käufer haben unterschiedliche Prioritäten, etwa Budget, Umweltfreundlichkeit oder technische Details, die der Chatbot erkennen und in personalisierte Antworten umsetzen muss.
Die Verarbeitung großer Datenmengen und die Integration aktueller Marktdaten in Echtzeit sind komplex. Unvollständige oder vage Anfragen erfordern gezielte Rückfragen. Zudem sind automatische Metriken wie BLEU oder ROUGE für die Bewertung der Beratungsqualität nur bedingt aussagekräftig, weshalb eine geeignete Infrastruktur für Human Feedback notwendig ist. Schließlich muss die Technologie eine intuitive Benutzeroberfläche bieten, um eine positive Nutzererfahrung sicherzustellen. Diese Herausforderungen machen fortschrittliche KI-Methoden und eine sorgfältige Planung unverzichtbar.

Entwicklung und Optimierung des Fahrzeugkaufberatungs-Chatbots

Der Fahrzeugkaufberatungs-Chatbot basiert auf fortschrittlichen Large Language Models (LLMs) wie LLAMA 3.2 [6] die mithilfe transformerbasierter Architekturen komplexe Muster in umfangreichen Textdatensätzen erkennen. Diese vortrainierten Modelle sind vielseitig einsetzbar, etwa zur Beantwortung von Fragen oder zur Analyse technischer Daten. Ziel ist es, das Modell speziell für die Fahrzeugkaufberatung zu optimieren, um präzise und relevante Antworten zu technischen Spezifikationen, Markttrends und individuellen Nutzerpräferenzen zu liefern.
Um diese Spezialisierung zu erreichen, ist ein Fine-Tuning des Modells auf domänenspezifische Datensätze geplant. Diese enthalten Informationen zu Fahrzeugmodellen, technischen Eigenschaften und Marktanalysen. Mithilfe von Low-Rank Adaptation [3] (LoRA) werden gezielt nur ausgewählte Parameter des Modells optimiert, was den Speicher- und Rechenbedarf erheblich reduziert. Ein weiterer geplanter Schritt ist die Integration eines Retrieval-Augmented Generation (RAG)-Moduls, das externe Wissensdatenbanken in Echtzeit durchsucht. Damit kann der Chatbot neben vortrainiertem Wissen auch dynamisch aktuelle Informationen wie Marktpreise und technische Details einbinden.
Ein zentraler Bestandteil der Optimierung ist die Nutzung von Reinforcement Learning from Human Feedback (RLHF). Hierbei werden Nutzerbewertungen, wie Upvotes und Downvotes, als Feedbacksignale verwendet, die durch Proximal Policy Optimization [4] (PPO) iterativ in die Verbesserung des Modells einfließen. Dieser Ansatz stellt sicher, dass hilfreiche Antworten verstärkt und weniger hilfreiche minimiert werden. Durch die Kombination von Fine-Tuning, RAG und kontinuierlichem Nutzerfeedback wird die Qualität der generierten Antworten nachhaltig verbessert. Der Kommunikationsablauf des Chatbots wird im folgenden Diagramm veranschaulicht. Der Workflow dieses Chatbots ist in Abbildung 1 visualisiert.

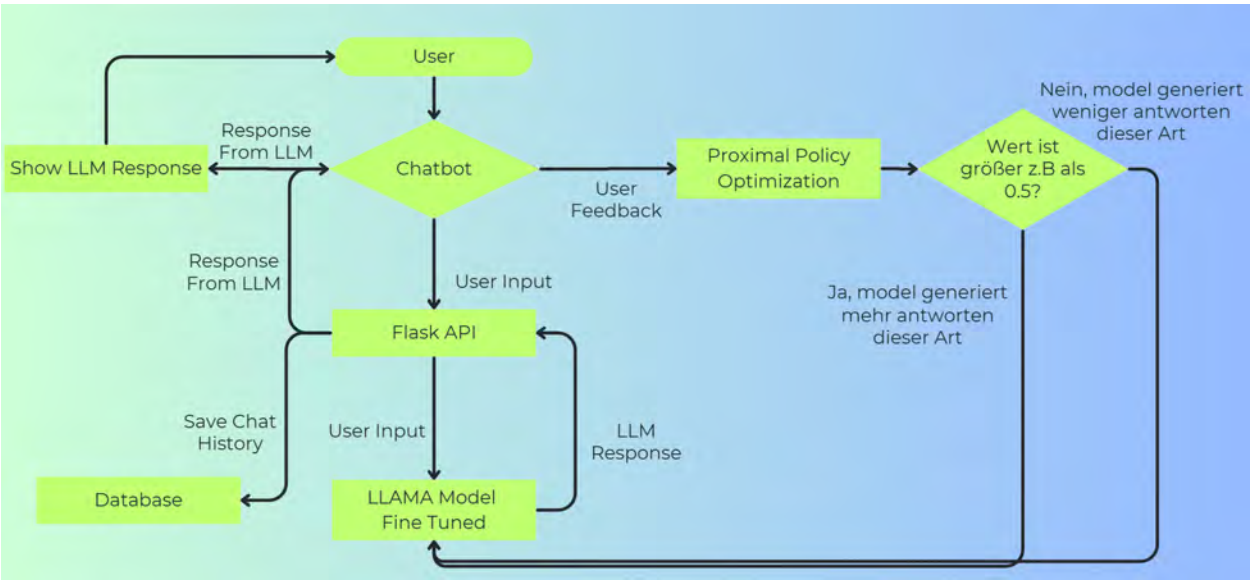


Abb. 1: Workflow Chatbot [2]

Nutzerinteraktion und Funktionsweise des Fahrzeugkaufberatungs-Chatbots

Der Fahrzeugkaufberatungs-Chatbot wurde entwickelt, um eine intuitive und interaktive Benutzererfahrung zu bieten. Die Benutzeroberfläche ermöglicht sowohl Texteingaben als auch das Hochladen von Bildern. Diese Kombination aus Text- und Bildverarbeitung schafft eine vielseitige Beratungsumgebung, die über einfache Textanfragen hinausgeht und auf spezifische Bedürfnisse der Nutzer eingeht.
Der Ablauf der Nutzerinteraktion gestaltet sich wie folgt:

- Texteingabe: Nutzer können Fragen oder Anforderungen eingeben, wie beispielsweise: „Welches Auto ist für Familien geeignet?“. Die Texteingabe wird über die Benutzeroberfläche an das Backend weitergeleitet, wo das LLM (Large Language Model) die Anfrage verarbeitet und eine präzise Antwort generiert.
- Bildbasierte Beratung: Neben der Texteingabe können Nutzer Bilder von Fahrzeugen hochladen, um beispielsweise ein Automodell zu identifizieren oder herauszufinden, ob es ihren Anforderungen entspricht. Ein Bildverarbeitungsmodul – basierend auf einem Bilderkennungsalgorithmus oder einem vortrainierten Modell – analysiert das hochgeladene Bild, erkennt das Fahrzeugmodell und konvertiert die Ergebnisse in einen Prompt, der an das LLM weitergegeben wird.
- Antwortgenerierung: Das LLM verarbeitet die Eingabe (Text oder Bild) und kombiniert vortrainiertes Wissen, branchenspezifische Informationen und, wenn verfügbar, aktuelle Daten aus dem Retrieval-Augmented Generation (RAG)-Modul. Die generierte Antwort wird dem Nutzer direkt auf der Benutzeroberfläche angezeigt.
- Feedback: Nutzer können die Qualität der Antwort durch Upvotes oder Downvotes bewerten. Diese Bewertungen werden in einer Datenbank gespeichert und dienen als Feedbacksignal für die spätere Optimierung des Modells mithilfe von Proximal Policy Optimization (PPO). Dieser Prozess stellt sicher, dass das Modell kontinuierlich durch Nutzerfeedback verbessert wird.

Der nahtlose Übergang zwischen Texteingabe, Bildverarbeitung und Feedback macht den Fahrzeugkaufberatungs-Chatbot zu einem vielseitigen und benutzerfreundlichen Tool. Die Integration moderner KI-Technologien sorgt dafür, dass Nutzer eine personalisierte und effiziente Beratung erhalten, die auf ihre individuellen Anforderungen zugeschnitten ist. Die Nutzerinteraktion beim Hochladen eines Bildes mit dem vortrainierten LLAMA 3.2 ist in Abbildung 2 dargestellt.

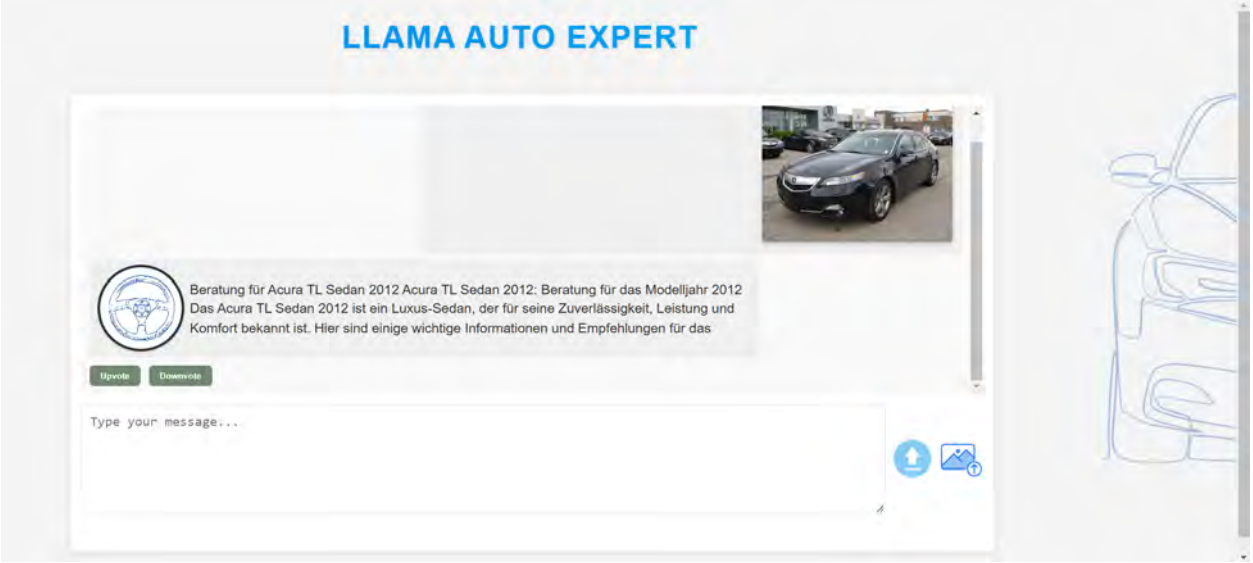


Abb. 2: Vortrainiertes LLAMA 3.2 3B Integriert im Chatbot [2]

Erwartete Ergebnisse und Ausblick

Es wird erwartet, dass der Chatbot präzise und personalisierte Antworten liefert, die sowohl technische Spezifikationen als auch individuelle Präferenzen der Nutzer berücksichtigen. Durch kontinuierliche Optimierung mithilfe von Nutzerfeedback wird der Chatbot langfristig flexibler und besser in der Lage sein, sich an wechselnde Anforderungen anzupassen. Zudem sollen die BLEU- und ROUGE-Werte im Vergleich zum vortrainierten Modell signifikant verbessert werden, was die Qualität der generierten Antworten bestätigt.
Zukünftig könnte der Chatbot um Multimodalität erweitert werden, um auch Sprache zu verarbeiten. Dies würde die Nutzererfahrung weiter optimieren und neue Anwendungsbereiche eröffnen, etwa für die sprachgesteuerte Interaktionen.
Darüber hinaus bietet die Skalierbarkeit des Systems Potenzial für Verbesserungen. Der Einsatz effizienterer Modellvarianten und serverloser Architekturen könnte den Ressourcenbedarf senken und die Verfügbarkeit des Systems erhöhen. Langfristige Nutzertests und eine kontinuierliche Verbesserung der Benutzeroberfläche bleiben entscheidend, um die Nutzerakzeptanz zu steigern und die Beratungsqualität weiter zu erhöhen. Der Chatbot dient als Proof of Concept und soll erste Einblicke in die Möglichkeiten einer automatisierten Fahrzeugkaufberatung bieten.

Literatur und Abbildungen

[1] European Automobile Manufacturers ACEA. Artificial Intelligence in the automobile industry. https://www.acea.auto/files/ACEA_Position_Paper-Artificial_Intelligence_in_the_automotive_industry.pdf, 11 2020.

[2] Eigene Darstellung.

[3] Hu Edward et al. LoRA: Low-Rank Adaptation of Large Language Models. https://arxiv.org/pdf/2106.09685, 2021.

[4] Schulman John et al. Proximal Policy Optimization Algorithms. https://arxiv.org/pdf/1707.06347, 2017.

[5] Gupta Mehul. LLM Evaluation metrics explained. https://medium.com/data-science-in-your-pocket/llm-evaluation-metrics-explained-af14f26536d2, 06 2024.

[6] Inc. Meta Platforms. Llama 3.2. https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_2, 2024.

[7] Gaoa Yunfan et al. Retrieval-Augmented Generation for Large Language Models: A Survey. https://arxiv.org/pdf/2312.10997, 12 2023.

Gaussian Splatting: Eine effiziente und skalierbare Methode zur 3D-Szenendarstellung

Alperen Akkurt **Dieter Morgenroth**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

Die bildbasierte 3D-Szenerekonstruktion mit Punkt basiertem Rendern (PBR) hat sich von frühen Ansätzen bis zu modernen Technologien wie NeRF und 3D Gaussian Splatting (GS) erheblich weiterentwickelt. Während frühere Methoden durch dichte Abtastung und strukturierte Aufnahmen eingeschränkt waren [1], berechnet Structure-from-Motion (SfM) aus einer Serie von 2D-Bildern die Kamerapositionen und eine spärliche 3D-Geometrie der Szene. SfM verwendet Merkmalsextraktion und -abgleich, um korrespondierende Punkte in mehreren Ansichten zu identifizieren und ihre räumliche Position durch geometrische Optimierung zu bestimmen [10]. Neural Radiance Fields erzielte einen Durchbruch in PBR, indem es mithilfe neuronaler Netze räumliche Koordinaten direkt in Farbe und Dichte umwandelte. Diese Methode hatte jedoch zwei wesentliche Nachteile, eine hohe Rechenintensität und eine begrenzte Editierbarkeit der Szenen [9]. 3D GS wurde als bahnbrechender neuer Ansatz entwickelt, der diese Einschränkungen überwindet. GS modelliert Szenen durch Millionen lernbarer splats im Raum, ermöglicht Echtzeit-Rendering ohne erheblichen Qualitätsverlust, bietet eine bessere Editierbarkeit durch explizite Szenenrepräsentation, somit vermeidet rechenintensive Berechnungen [8]. Diese Innovation eröffnet neue Möglichkeiten in Bereichen wie Virtual/Augmented Reality und autonomes Fahren. Die schnell wachsende Bedeutung von 3D GS zeigt sich auch in der steigenden Anzahl an wissenschaftlicher Publikationen seit seiner Einführung (siehe Abbildung 1) [5].

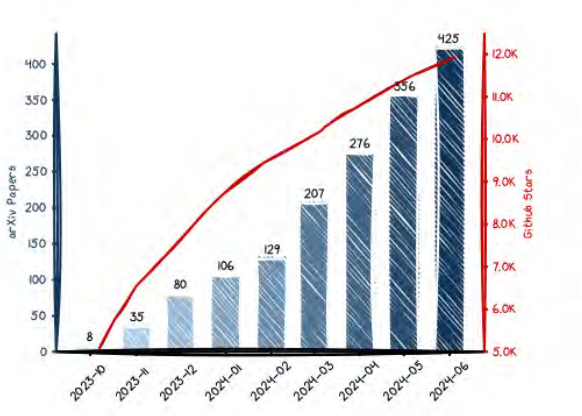


Abb. 1: Publikationen und GitHub-Sterne für 3D GS [5]

Motivation und Problemstellung

Die Darstellung und Verarbeitung von 3D-Szenen ist ein grundlegendes Problem vieler wissenschaftlicher und technischen Disziplinen. Von der medizinischen Bildgebung über die Simulation physikalischer Prozesse bis hin zur Entwicklung autonomer Systeme spielt die effiziente und akkurate Repräsentation von 3D-Daten eine entscheidende Rolle. Traditionelle Methoden, wie beispielsweise polygonale Meshes, stoßen bei der Darstellung komplexer Szenen mit hoher Detailtreue an ihre Grenzen. Die zunehmende Verfügbarkeit von 3D-Sensoren und die steigende Nachfrage nach immersiven Erlebnissen, wie sie in Virtual und Augmented Reality Anwendungen gefordert werden, verschärfen diese Problematik zusätzlich [7]. Das explosionsartige Wachstum der von 3D-Scannern erfassten Datenmengen stellt eine Herausforderung für Verarbeitung, Speicherung und Visualisierung dar und erfordert neue, effizientere Methoden zur 3D-Szenendarstellung. Gaussian Splatting bietet einen vielversprechenden Ansatz, um diesen Herausforderungen zu begegnen.

Forschungsfrage und eigener Beitrag: Diese Arbeit zielt darauf ab, die Einsatzmöglichkeiten von Gaussian Splatting bei der Erstellung hochdetaillierter 3D-Umgebungen aus Kameraaufnahmen zu untersuchen und dessen Integration in realistische Anwendungsumgebungen zu erforschen. Im Fokus steht die Optimierung des Verfahrens für Kameraaufnahmen und die nahtlose Einbindung der erzeugten 3D-Szenen in Unreal Engine. Dabei wird eine hochdetaillierte 3D-Karte generiert und mit einem bereitgestellten 3D-Fahrzeugmodell kombiniert, um ein realistisches Fahrerlebnis in einer simulierten Umgebung zu schaffen.

Stand der Forschung

Die Entwicklung der 3D-Szenendarstellung hat sich von traditionellen geometriebasierten Ansätzen hin zu modernen Deep-Learning-Methoden erheblich weiterentwickelt. Frühe Ansätze wie BundleFusion basierten auf geometrischen Rekonstruktionen, die durch mathematische Modelle präzise Szenen abbildeten. Diese Methoden waren jedoch durch hohe Anforderungen an Rechenleistung und Speicher sowie die eingeschränkte Darstellung komplexer Szenen limitiert [2]. Punktbasierte Darstellungen, bei denen Szenen durch diskrete Punkte anstelle von Polygonen repräsentiert wurden, boten eine flexible Alternative für unstrukturierte Daten, hatten aber mit Problemen wie Löchern und Aliasing zu kämpfen. Volumetrische Methoden erweiterten die Darstellungsmöglichkeiten, indem sie Objekte und Phänomene wie Rauch oder Nebel als Volumen modellierten, waren jedoch durch die Rechenintensität des Ray-Marchings begrenzt. Neural Radiance Fields (NeRF) brachten einen Durchbruch durch die Nutzung neuronaler Netze für fotorealistische Rekonstruktionen mittels Positionskodierung und volumetrischem Ray-Marching. NeRF überzeugte durch hohe Bildqualität, blieb aber in der Geschwindigkeit und Editierbarkeit eingeschränkt [9]. Gaussian Splatting kombiniert die Stärken früherer Ansätze mit Deep Learning und ermöglicht effizienteres, skalierbares und editierbares 3D-Rendering. Der besondere Vorteil liegt in der Differenzierbarkeit der 3D-Gaussians, was die einfachere Optimierung ermöglicht.

Gaussian Splatting Pipeline

Die Gaussian-Splatting Pipeline besteht aus mehreren Phasen

Datenerfassung und Vorverarbeitung: Zunächst werden hochwertige Bild- oder Videodaten erfasst. Bei Videos erfolgt die Frameextraktion mit FFmpeg [11]. Essenziell sind verschiedene, überlappende Blickwinkel für eine umfassende Szenendarstellung.Die Vorverarbeitung umfasst die Vereinheitlichung der Auflösung, Farbkorrekturen zur Normalisierung der Beleuchtung und die Entzerrung von Linsenverzerrungen.

Geometrische Prioren: Die präzise Bestimmung der geometrischen Prioren bildet das Fundament der 3D-Rekonstruktion. Structure-from-Motion Algorithmen, implementiert in COLMAP, generieren eine initiale spärliche Punktwolke und ermitteln die ungefähren Kamerapositionen im dreidimensionalen Raum. Diese Initialisierung ist von fundamentaler Bedeutung, da ohne akkurate Kameraparameter eine korrekte 3D-Rekonstruktion mathematisch nicht möglich wäre [10].

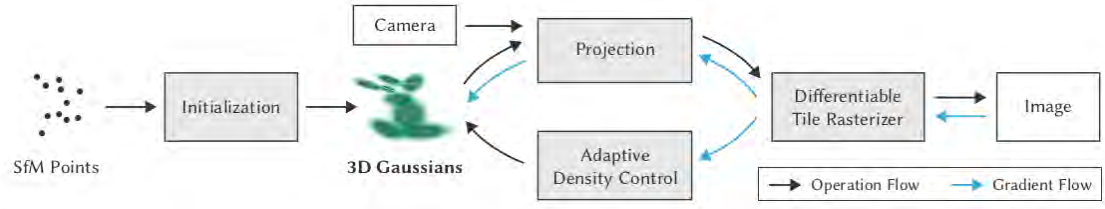


Abb. 2: Gaussian Pipeline [8]

Splat-Initialisierung: Nun folgt der eigentliche Einstieg in Gaussian Splatting mit der Initialisierung der Splat Parameter (siehe Abbildung 2, die unten beschrieben Parameter werden für jeden Splat einzeln festgelegt.

- Position des Splat-Mittelpunkts: $\mu \in R^3$
- Kovarianzmatrix: $\Sigma \in R^{3\times3}$, die sich zerlegen lässt als $\Sigma = RSS^\top R^\top$ wobei R die Rotationsmatrix und S die Skalierungsmatrix ist.
- Opazität $\alpha \in [0, 1]$
- Erscheinungsbild $c \in R^3$.

Die Position μ kann anhand der Punktwolke bestimmt werden. Die Kovarianzmatrix beschreibt die Form und Orientierung, wie angeben oft durch Rotation und Skalierung parametrisiert. Die Opazität steuert die Transparenz, und das Erscheinungsbild beinhaltet Farbinformationen und gegebenenfalls sphärische Harmonisch [8].

Splat Rendern: Die 3D-Gaussians werden auf die 2D-Bildebene projiziert, wobei ein differenzierbarer

Rasterisierer die Pixelabdeckung und den Farbbeitrag jedes Splats effizient berechnet, dies lässt sich in drei wesentliche Transformationsschritte unterteilen:
1. Kamera-Projektion

$$\mathbf{x}' = \mathbf{P}(\mathbf{R}\mu + \mathbf{t}) \quad (1)$$

In 1 wird die Transformation der 3D-Positionen der Gaussians in den 2D-Bildraum, indem zunächst eine Rotation R und Translation t die Position in das Kamerakoordinatensystem überführt und anschließend die Projektionsmatrix P diese in den 2D-Bildraum projiziert [8].
2. Kovarianz-Transformation

$$\Sigma' = \mathbf{JWW}^\top \mathbf{J}^\top \quad (2)$$

Die Kovarianzmatrix wird durch angebene Formel in den Bildraum transformiert. Dabei wird die Jacobi-Matrix J der Projektion verwendet, welche die ursprüngliche 3D-Kovarianz W berücksichtigt und in einer 2D-Kovarianzmatrix Σ' resultiert, die die Form des projizierten Splats beschreibt [8].
3. Alpha-Compositing

$$C = \sum_{i=1}^{N} \mathbf{c}_i \alpha_i' \prod_{j=1}^{i-1} (1 - \alpha_j') \quad (3)$$

Die Farbberechnung erfolgt durch Alpha-Compositing, bei dem überlappende Splats entsprechend ihrer Tiefenreihenfolge miteinander vermischt werden. Dabei werden die Farbbeiträge aller Splats mit ihrer jeweiligen Opazität gewichtet und die Verdeckung durch davor liegende Splats einberechnet, um eine differenzierbare Berechnung der finalen Pixelfarbe zu gewährleisten [8].
Verlustberechnung und Optimierung: Die Verlustberechnung und -optimierung verfeinern die Splat-Parameter iterativ, um die Differenz zwischen gerenderten und Eingabedaten zu minimieren. Verwendet werden Verlustfunktionen wie L1-Verlust, der die absolute Differenz misst, und D-SSIM (Structural Similarity Index Measure), der die strukturelle Ähnlichkeit bewertet. Die Gesamtverlustfunktion kann eine gewichtete Kombination dieser Metriken sein. Die Optimierung nutzt gradientenbasierte Methoden wie SGD oder Adam, wobei die differenzierbare Rasterisierung die effiziente Gradientenberechnung ermöglicht. Die Lernrate wird angepasst, um eine optimale Konvergenz zu erreichen.
Die Gesamtverlustfunktion wird wie folgt berechnet:

$$L = (1 - \lambda)L_1 + \lambda L_{D-SSIM} \quad (4)$$

Um eine optimale Balance zwischen Genauigkeit und Rechenzeit zu finden, wird das Splat-Rendering iterativ durchgeführt, bis ein definiertes Abbruchkriterium (z.B. Konvergenz der Verlustfunktion oder maximale Iterationszahl) erreicht ist (siehe Abbildung 2) [8].

Ausgabe: Das Endresultat des Gaussian Splatting ist eine kompakte Szenenrepräsentation, die aus optimierten Gaussian-Splats besteht. Die Visualisierung kann auf verschiedenen Plattformen erfolgen. Dank der Integration mit PlayCanvas und SuperSplat ist auch eine direkte Präsentation im Webbrowser möglich, ohne zusätzliche Software zu installiert [4].

Implementierung

Abb. 3: Calzifer Plüschfigur [6]

Die gezeigte Stoffpuppe in Abbildung 3 zeigt eine Plüschfigur mit feinen Details wie Härchen, der charakteristische Farbverlauf von orange zu rot sowie die organische Form, dies demonstriert die besonderen Herausforderungen bei der 3D-Rekonstruktion komplexer Objekte.stellen für traditionelle 3D-Modellierungstechniken eine besondere Herausforderung dar. Dies ist für konventionelle Computergrafik, die auf Polygonen und Meshes basiert mit erheblichem Zeitaufwand in Verbindung, Gaussian Splatting bietet hier eine effiziente Alternative zur detailgetreuen Rekonstruktion
Die dreidimensionale Rekonstruktion erfolgt durch einfache einminütige Videoaufnahme bspw. mit einer Handykamera. Nach einer Trainingszeit von etwa 50 Minuten entsteht ein editierbares 3D-Modell (siehe Abbildung 4). Die Qualität der Rekonstruktion zeigt sich besonders in der detaillierten Darstellung der Stoffpuppe. Die ursprüngliche Rekonstruktion mit 100.000 Splats. können durch entfernen von Splats mit geringer Opazität (unter 0,2) auf 80.000 optimiert, ohne die visuelle Qualität zu beeinträchtigen.

Abb. 4: Calzifer 3D Rekonstruktion [3]

Die Abbildung zeigt drei verschiedene Darstellungen der rekonstruierten Plüschfigur: Das linke Bild präsentiert alle 80.000 Splats in ihrer Gesamtheit, während die mittlere und rechte Darstellung jeweils unterschiedliche Teilmengen von 40.000 Splats zeigen. Besonders bemerkenswert ist die Verteilung kleiner Splats, die für die hochwertige Darstellung der feinen Härchen der Stoffpuppe verantwortlich sind und somit zur realistischen Gesamterscheinung beitragen.

Ausblick

3D Gaussian Splatting (3D GS) bietet vielseitige Anwendungsmöglichkeiten und adressiert zentrale Herausforderungen in der Darstellung und Verarbeitung komplexer 3D-Szenen. In der Robotik, speziell in der Simultanen Lokalisierung und Kartierung (SLAM), ermöglicht 3D GS die präzise Echtzeitdarstellung von Umgebungen und erleichtert die Navigation durch dynamische Elemente. Für die Rekonstruktion dynamischer Szenen bietet 3D GS die Möglichkeit, zeitliche Veränderungen und Bewegungen genau zu modellieren, was besonders in der virtuellen Realität und Animation von Vorteil ist. In der kreativen Gestaltung revolutioniert 3D GS die Erstellung KI-generierter Inhalte, indem es hochrealistische und editierbare 3D-Assets in Echtzeit ermöglicht. Für autonomes Fahren hilft 3D GS, sensorbasierte Daten in kohärente Darstellungen umzuwandeln, um sowohl statische als auch dynamische Szenenelemente zuverlässig zu rekonstruieren. Schließlich kann 3D GS in der medizinischen Bildgebung und bei großskaligen Szenenrekonstruktionen effizient große Datenmengen verarbeiten, was es zu einem wertvollen Werkzeug in der Medizintechnik und Stadtmodellierung macht [5].

Literatur und Abbildungen

[1] Brian Curless and Marc Levoy. A Volumetric Method for Building Complex Models from Range Images. In *SIGGRAPH '96*. Stanford University, 1996.

[2] Angela Dai, Matthias Niessner, Michael Zollöfer, Shahram Izadi, and Christian Theobalt. BundleFusion: Real-time Globally Consistent 3D Reconstruction using On-the-fly Surface Re-integration. *ACM Transactions on Graphics 2017*, 2017.

[3] Eigene Darstellung.

[4] Hutchence Donovan. SuperSplat - 3D Gaussian Splat Editor. https://github.com/playcanvas/supersplat, 2024.

[5] Ben Fei, Jingyi Xu, Rui Zhang, Qingyuan Zhou, Weidong Yang, and Ying He. 3D Gaussian Splatting as New Era: A Survey. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–20, 2024.

[6] miyazaki hayao. L. https://www.donguri-sora.com/category/PLUSHDOLL/21502540.html, 2023.

[7] Furion analytics industryarc. 3D Sensors Market Overview. https://www.industryarc.com/Report/244/global-3D-sensor-market-analysis-report.html, 2024.

[8] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 2023.

[9] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *Computer Vision – ECCV 2020*, pages 405–421. Springer International Publishing, 2020.

[10] Johannes L. Schonberger and Jan-Michael Frahm. Structure-From-Motion Revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Computer Vision and Pattern Recognition (CVPR), 2016.

[11] Suramya Tomar. Converting video formats with FFmpeg. *Linux Journal*, 2006.

V2X-basierte CACC-Entwicklung: Entwicklung und Integration eines V2X-Datenfusionsmodells sowie die Analyse von Bremslogiken in Mischverkehrsszenarien

Felix Anslinger **Reiner Marchthaler**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Porsche Engineering Services GmbH, Mönsheim

Einführung

Die fortschreitende Digitalisierung prägt den modernen Straßenverkehr und eröffnet neue Möglichkeiten zur Verbesserung von Sicherheit, Effizienz und Komfort. Dabei spielen Fahrerassistenzsysteme eine zentrale Rolle, indem sie den Fahrer aktiv unterstützen und so das Risiko von Unfällen reduzieren sowie den Fahrkomfort steigern. [3] Mit den Fortschritten in Sensorik, Algorithmen und Kommunikationssystemen werden diese Systeme stetig leistungsfähiger und vielseitiger. Besonders die Kommunikation zwischen Verkehrsteilnehmern, ermöglicht durch Technologien wie Vehicle-to-Everything (V2X), hat in den letzten Jahren erheblich an Bedeutung gewonnen und schafft die Grundlage für innovative Ansätze im kooperativen Fahren. [2] Im Rahmen der vorliegenden Arbeit wird solch ein Ansatz entwickelt, der darauf abzielt, die Potenziale der modernen Fahrzeugkommunikation zur Erweiterung von bereits bestehenden Fahrerassistenzsystemen zu nutzen.

Grundlagen der V2X-Technologie

Der Begriff "Vehicle-to-Everything"(V2X) bezeichnet die Kommunikation zwischen Fahrzeugen und ihrer Umgebung. Ziel ist es, durch V2X die Sicherheit, Effizienz und den Komfort im Straßenverkehr zu erhöhen. Als zentraler Bestandteil intelligenter Transportsysteme (ITS) ermöglicht V2X eine vernetzte Zusammenarbeit aller Verkehrsteilnehmer durch den Austausch relevanter Daten. Hierbei wird die V2X-Kommunikation in verschiedene Typen unterteilt, wie Vehicle-to-Vehicle (V2V), Vehicle-to-Infrastructure (V2I), Vehicle-to-Pedestrian (V2P), Vehicle-to-Network (V2N), Vehicle-to-Grid (V2G) und Vehicle-to-Device (V2D), je nachdem, welche Verkehrsteilnehmer miteinander kommunizieren (siehe Abbildung 1). [4] Der Datenaustausch erfolgt dabei über standardisierte Nachrichten wie Cooperative Awareness Messages (CAM), Collective Perception Messages (CPM) und Decentralized Environmental Notification Messages (DENM), die jeweils auf spezifische Informationen ausgelegt sind. Die Technologie leistet somit einen maßgeblichen Beitrag zur Erhöhung der Verkehrssicherheit, zur Optimierung des Verkehrsflusses sowie zur effizienten Integration autonomer Fahrfunktionen.

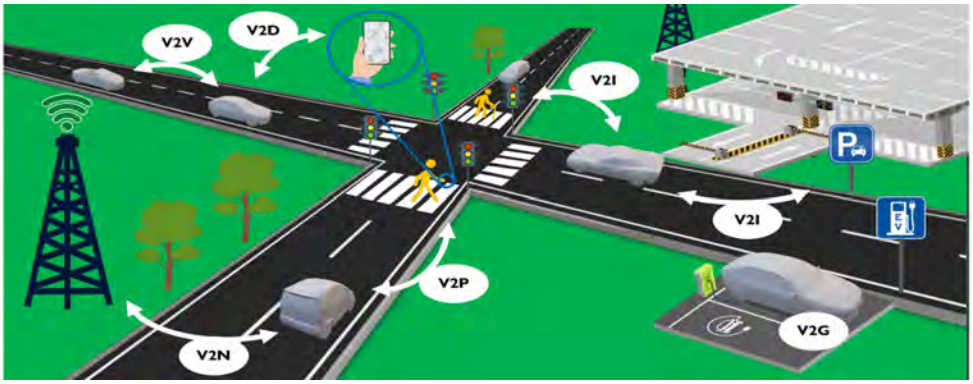


Abb. 1: Verschiedene V2X-Kommunikationstypen. [5]

Erweiterung von ACC zu CACC

Adaptive Cruise Control (ACC) bezeichnet ein Fahrerassistenzsystem, welches den Abstand zu vorausfahrenden Fahrzeugen durch automatische Geschwindigkeitsanpassung reguliert. Dadurch werden sowohl der Komfort als auch die Sicherheit im Straßenverkehr erhöht. [3] Im Gegensatz zum herkömmlichen Tempomaten reagiert ACC dynamisch auf Verkehrssituationen, indem es mithilfe von Fahrzeugsensoren Daten wie Distanz und Geschwindigkeit erfasst. Die Verwendung von Sensordaten ermöglicht eine verlässliche Abstandskontrolle, auch bei wechselnden Bedingungen. Allerdings ist die Funktionalität von ACC durch die Reichweite und Zuverlässigkeit der verwendeten Sensoren begrenzt.
Eine Weiterentwicklung des zuvor beschriebenen Systems stellt die Cooperative Adaptive Cruise Control (CACC) dar, bei der das ACC durch die Integration von V2X-Kommunikation, insbesondere Vehicle-to-Vehicle (V2V), ergänzt wird. Der direkte Informationsaustausch zwischen Fahrzeugen ermöglicht es CACC, nicht nur präziser und schneller auf das Verhalten anderer Verkehrsteilnehmer zu reagieren, sondern auch das Verhalten weiter entfernter Fahrzeuge in einer Kolonne zu antizipieren (siehe Abbildung 2). Dies führt zu einer verbesserten Verkehrssicherheit, da die Systemreaktionszeiten deutlich reduziert werden und Sensorgrenzen, beispielsweise durch widrige Wetterbedingungen, durch V2X-Daten kompensiert werden können. Trotz dieser Vorteile stellt die noch geringe Verbreitung der V2X-Technologie im Fahrzeugmarkt eine Herausforderung dar, da die Effektivität von CACC von der Anzahl der vernetzten Fahrzeuge abhängt. Um diesen Übergang bewältigen zu können, müssen CACC-Systeme so konzipiert sein, dass sie auch in Mischszenarien von vernetzten und nicht vernetzten Fahrzeugen sicher und zuverlässig funktionieren.

Abb. 2: Beispielhafter Informationsfluss bei CACC. [1]

Zentrales Forschungsziel der Arbeit

Das zentrale Forschungsziel dieser Arbeit ist die Entwicklung eines robusten CACC-Systems, welches den Herausforderungen des Mischverkehrs mit unterschiedlich ausgestatteten Fahrzeugen gerecht wird. Dabei soll ein Ansatz entwickelt werden, der V2X-Daten aus verschiedenen Nachrichtentypen wie CAM, CPM und DENM effizient fusioniert. Das System soll demnach nicht nur auf vernetzte Fahrzeuge reagieren, sondern auch Informationen über nicht vernetzte Fahrzeuge, welche von vernetzten Fahrzeugen erfasst und weitergeleitet werden, einbeziehen. Ein weiterer Fokus liegt auf der Konzeption und Evaluierung diverser Algorithmen für die CACC-Bremslogik. Hierbei wird untersucht, welche Bremsstrategien, unter Berücksichtigung der erhaltenen V2X-Informationen, sich am besten zur Optimierung von Sicherheit und Fahrkomfort eignen.

Literatur und Abbildungen

[1] CAR TO CAR Communication Consortium. C2C-CC Basic System Functionality and Use Cases. https://www.car-2-car.org/fileadmin/documents/General_Documents/C2CCC_UC_2097_UseCases_V1.0.pdf, 2023.

[2] Bethan Grylls. More than 11.2m vehicles will have V2X communications in 2024. *New Electronics*, 2019.

[3] Ye Li, Zhibin Li, Hao Wang, Wei Wang, and Lu Xing. Evaluating the safety impact of adaptive cruise control in traffic oscillations on freeways. *Accident Analysis & Prevention*, 2017.

[4] Ignacio Soto, Maria Caldero, Oscar Amador, and Manuel Urueña. A survey on road safety and traffic efficiency vehicular applications based on C-V2X technologies. *Vehicular Communications*, 2022.

[5] Vygantas Ušinskis, Mantas Makulavičius, Sigitas Petkevičius, Andrius Dzedzickis, and Vytautas Bučinskas. Towards Autonomous Driving: Technologies and Data for Vehicles-to-Everything Communication. *Sensors*, 2024.

Konzeptionierung und Realisierung eines verteilten, echtzeitfähigen, CAN-basierten Firmware-Update-Tools

Manuel Athanasas **Walter Lindermeir**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma comemso electronics GmbH, Ostfildern

Einleitung

Die Automobilindustrie befindet sich in letzter Zeit in immer stärker werdenden Unruhen. Immer mehr Unternehmen entlassen massenhaft Personal oder gehen sogar in die Insolvenz. Daher ist es notwendig Kunden den Zugang zu neuen Funktionalitäten eines einmal gekauften Produkts so einfach wie möglich zugänglich zu machen. Gerade die Elektromobilität befindet sich noch in der Forschung, weshalb es auch stetig Neuerungen und Verbesserungen gibt. Diese sind oft von Normungsinstitutionen wie der DIN oder ISO initiiert. Aus diesem Grund sollten Erweiterungen und Fehlerbehebungen von Funktionalitäten in einem System bestmöglich auf dem Gerät möglich sein. Klassischerweise findet dies in Form eines Firmwareupdates statt. Ein solches Update genügt oft, solange keine neue Hardware nötig ist. Aufgrund von weltweitem Einsatz der Geräte muss die Frage gestellt werden, wie ein Firmwareupdate durchgeführt wird. Es gibt hierfür mehrere Möglichkeiten: Entweder muss ein geschulter Mitarbeiter des Herstellers zum Verwendungsort des Geräts reisen oder das Gerät kommt zu ihm. Weil beides allerdings wenig praktikabel ist, gibt es einen weiteren Ansatz. Der Endanwender kann auch selbstständig ein Update durchführen, wenn der Update-Prozess weitestgehend automatisiert abläuft. Dadurch wird der Aufwand des Herstellers im gleichen Schritt minimiert.

Aufgabenstellung

Die Aufgabenstellung dieser Arbeit ist es den vorhandenen Firmwareupdateprozess zu optimieren. Gegenwärtig wird dieser Prozess komponentenweise durchgeführt. Ein System besteht im Regelfall aus mehreren Komponenten. In Zukunft soll ein Update für das gesamte System durchgeführt werden und nicht für einzelne Komponenten. Der Anwender kann somit den Update-Prozess anstoßen und dieser läuft ohne weitere Interaktionen automatisiert ab.

Bootloader

Ein Bootloader ist eine Software, die beim Start eines Systems die Applikation auf einen Mikrocontroller lädt. Diese Funktionalität ist extrem kritisch, da bei einem Fehler der Ausfall des Systems resultieren kann. Im Update-Fall wird die Applikation von einem Hostprogramm an einem PC über eine vordefinierte Schnittstelle an den Mikrocontroller übertragen. Diese Schnittstelle ist standardmäßig nicht die bereitgestellte Verbindungsart des Mikrocontrollers wie beispielsweise die JTAG-Schnittstelle. Der Unterschied ist, dass der Entwickler die Schnittstelle selbst implementieren muss, so aber Zusatzinformationen während des Updates durch die Eigenimplementierung mitgeben kann. Von Vorteil ist die Realisierung eines Bootloaders, da so noch weitere Funktionen umgesetzt werden können, die ein Update vereinfachen, wie beispielsweise ein Versionsvergleich der Applikation vor Durchführung eines Updates. Ein Bootloader wird vor allem im Einsatzbereich von eingebetteten Systemen benötigt, da dort der Mikrocontroller schlecht erreichbar ist. Um trotzdem Updates zu ermöglichen wird deshalb eine Schnittstelle definiert, die es ermöglicht an einem erreichbaren Ort mit dem Mikrocontroller zu kommunizieren. [1] Das kann vom Herausführen eines Steckers bis hin zu einem Firmwareupdate over the Air, kurz FOTA, reichen. Je nach Variante wird hierfür aber Hardware benötigt, die nicht im Mikrocontroller integriert ist. Ein naheliegendes Beispiel ist ein Steuergerät in einem Auto. Dies ist vom Endanwender nicht erreichbar, muss aber trotzdem Updates erhalten können. Deshalb bekommen zunehmend Anwender die Möglichkeit ein Firmwareupdate ihres Autos selbst durchzuführen. Eine weitere Möglichkeit, die ein Bootloader bietet, ist das parallele Speichern von zwei Applikationen in einem Mikrocontroller. Hierbei werden zwei Varianten einer auszuführenden Firmware für den Mikrocontroller in einen Flashspeicher geladen. Während der Bootloader ausgeführt wird, entscheidet dieser welche Applikation gestartet werden soll.

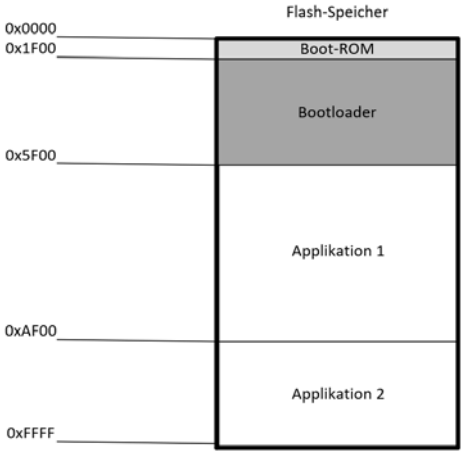


Abb. 1: Bootloader mit mehreren Applikationen [2]

Der Mikrocontroller führt seine grundlegenden Initialisierungen (Boot-ROM) aus. In der Logik des Bootloaders müssen die Minimalanforderungen programmiert werden, sodass auch eine Applikation auf diesem Mikrocontroller laufen kann. Diese Anforderungen sind das Laden einer Applikation in den Flash-Speicher des Mikrocontrollers und die unabhängige Funktionsweise von der Applikation. Der Bootloader darf keine Funktionalität aus der Applikation verwenden, da der Bootloader den Speicherbereich dieser während der Aktualisierung überschreibt. Die Applikation hingegen darf auf den Code des Bootloaders zugreifen. Der Bootloader sollte so klein wie möglich gehalten werden, da er im Speicher verbleibt und somit für die Applikation Speicher belegt. Um die Minimalanforderung zu erfüllen, muss sich aber noch über die Wahl der Schnittstelle Gedanken gemacht werden.

Peripheriemöglichkeiten

Normalerweise wird eine Punkt-zu-Punkt Verbindung verwendet. Das hat zur Folge, dass die Kommunikation von Mikrocontroller und Host isoliert ist und so von äußeren Störfaktoren, wie anderen Kommunikationsteilnehmern geschützt bleibt. Diese Lösung lässt immer einen Host mit genau einem Mikrocontroller kommunizieren. Somit kann auch nur ein Update gleichzeitig ausgeführt werden. Werden mehrere Updates in einem System nötig, muss nach dem Update einer Komponente die oben genannte Punkt-zu-Punkt Verbindung manuell umgesteckt werden. Der Ansatz eines Kommunikationsnetzes mit einer Switch-basierten Sterntopologie versucht den Nachteil einer solchen Punkt-zu-Punkt Verbindung zu umgehen. Der Host wird mit den Komponenten über eine Punk-zu-Multipunkt Verbindung verbunden. So ist es dem Host möglich mehrere Komponenten gleichzeitig anzusteuern. Bei dem gleichen Verfahren wie der Punkt-zu-Punkt Verbindung könnten so mehrere Updates nacheinander durchgeführt werden ohne die manuelle Veränderung einer Verbindung zu nötig zu machen. Um eine solche Sterntopologie zu ermöglichen ist zusätzliche Hardware notwendig, der Switch müsste im System ergänzt werden, sodass sich der Host dort einbinden kann. Eine dritte Möglichkeit für ein Kommunikationsnetz bietet die Bustopologie, auch so wird eine Punkt-zu-Mehrpunktverbindung ermöglicht. Im Gegensatz zu dem vorangegangenen Ansatz verwenden die Kommunikationsteilnehmer ein gemeinsames Medium und sind dadurch verbunden. Infolge dessen wird keine zusätzliche Hardware benötigt, es muss lediglich sichergestellt sein, dass alle Komponenten mit dem Bus verbunden sind. [3] Bei einem Bus muss beachtet werden, dass eine isolierte Kommunikation nicht von anderen Busteilnehmern gestört wird. Das wäre der Fall sobald der Host eine Komponente updaten will, eine andere Komponente aber auch reagiert und beginnt zu kommunizieren. Der Bootloader und der Host müssen dementsprechend robust programmiert sein, um Fehler aufgrund solcher Störungen zu vermeiden. Um den Updateprozess zu automatisieren wird die Entwicklung durchaus komplexer, im Gegenzug wird die Durchführung dadurch vereinfacht.
Die Peripherieentscheidung sollte am Anfang des Entwicklungsprozesses getroffen werden, da je nach Wahl bestimmte Prinzipien und Algorithmen erforderlich sind, wie beispielsweise die robuste Programmierung der Kommunikation in einer Punkt-zu-Mehrpunkt-Kommunikation. Auf Basis dieser Entscheidung kann dann das restliche System ausgelegt werden.

Literatur und Abbildungen

[1] Rüdiger R. Asche. *Embedded Controller Grundlagen und praktische Umsetzung für industrielle Anwendungen*. Springer Fachmedien Wiesbaden GmbH, 2016.

[2] Eigene Darstellung.

[3] Matthias Rausch. *Kommunikationssysteme im Automobil*. Carl Hanser Verlag, 2022.

Entwicklung eines webbasierten Backlog-Analyzer

Benjamin Baunach **Jörg Nitzsche**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma LF Consult, Stuttgart

Einleitung

In der heutigen Zeit, in der Unternehmen zunehmend auf digitale Prozesse und Datenanalysen angewiesen sind, spielen effektive Tools zur Überwachung und Optimierung von Systemressourcen eine zentrale Rolle. Business Itelligence (BI) und Reporting bieten hierfür Lösungen, um Informationen aus komplexen Datenmengen zu extrahieren, um fundierte Entscheidungen treffen zu können und um Problemen frühzeitig entgegenzuwirken. Eine umfassende Analyse und Auswertung von Unternehmensdaten sind entscheidend, um Produkte erfolgreich am Markt zu etablieren und langfristig wettbewerbsfähig zu bleiben. Besonders wichtig ist es, diese Analysen so zugänglich und verständlich zu gestalten, dass auch ohne tiefgreifendes Fachwissen schnell und effizient Maßnahmen ergriffen werden können, die zu nachhaltigen Lösungen und einer langfristigen Optimierung führen.

Problemstellung

Die 3Liter-PPS®-Software von LF-Consult unterstützt produzierende Unternehmen dabei, das Konzept "Produzieren im Takt"(PiT®) umzusetzen. Zur Überwachung der 3Liter-PPS ® Software auf den Kundensystemen, sendet diese täglich Systemrelevante Daten an einen dezidierten E-Mail-Receiver. Von hier aus, werden die Daten in einer Datenbank gespeichert, um von einem bestehenden Reporting-Tool validiert zu werden. Hierbei handelt es sich um Daten wie bspw. Festplattenspeicherkapazitäten, Lizenzlaufzeiten, Dienste Status, speziellen Events usw. Die Validierung dieser Daten erfolgt anhand vordefinierter Schwellwerte, die manuell in einem Backend festgelegt wurden. Ziel des Reportings ist es, LF-Consult dabei zu unterstützen, kritische Fehler auf den Kundensystemen frühzeitig zu erkennen, zu beheben und nachhaltige Lösungen zur Fehlerprävention zu entwickeln. Das bestehende Reporting-Tool weist erhebliche Schwächen in Usability und Performance auf. Fehlermeldungen werden zwar auf der Weboberfläche angezeigt, doch die Ursachen sind oft schwer nachvollziehbar. Zudem fehlt eine Auswertung der gesammelten Daten, was die Entwicklung nachhaltiger Lösungen erschwert

Zielsetzung

Ziel der Bachelorarbeit ist es, für LF-Consult einen nachhaltigen Mehrwert zu schaffen, indem ein neues Reporting-Tool entwickelt wird, das langfristig eine geringe Fehlerquote sicherstellt und somit die Kundenzufriedenheit erhöht. Dies beinhaltet: Technische Verbesserungen wie die Anpassung an moderne Technologien mit Springboot und Angular. Die Entwicklung von Modularität und Skalierbarkeit zur einfachen Anpassung an neue Anforderungen. Die Entwicklung eines flexiblen Metriksystems für standardisierte und benutzerdefinierte Metriken. Aber auch eine Anpassung der Benutzerfreundlichkeit wie die klare und transparente Darstellung der Metriken für eine bessere Nachvollziehbarkeit und eine intuitive Weboberfläche auf der Basis des 3Liter-PPS®-Styleguides. Und schließlich fortschrittliche Analysen zur genauen Analyse von Fehlern, um Ursachen und Muster zu erkennen.

Reporting und Monitoring

Reporting (Berichtswesen) bezeichnet das Sammeln, Analysieren und Darstellen von Daten, um fundierte Entscheidungen und Maßnahmen zu ermöglichen. Der Begriff Reporting wird dabei häufig in Zusammenhang mit Controlling (Überwachung) und Monitoring (Überprüfen) verwendet. Während Monitoring vor allem für die Überwachung und Alarmierung von Prozessen zuständig ist, dient das Controlling der Steuerung und gezielter Nutzung von Unternehmensdaten. Reporting dient somit nicht nur der Darstellung von Daten, sondern bildet im Wesentlichen Unternehmensprozesse und deren Ergebnisse in Form realer Vorgänge ab. Dadurch wird es zu einem wichtigen Analysewerkzeug, das der Entwicklung gezielter Maßnahmen und Strategien dient. [3] In IT-Unternehmen ist es von entscheidender Bedeutung, Transparenz und Kontrolle über Systeme zu gewährleisten, um Schwachstellen zu identifizieren, gezielte Maßnahmen zu ergreifen und

Probleme effizient zu lösen. Ein zentraler Bestandteil dieses Prozesses ist das System-Reporting, das auf der Definition von spezifischen Kennzahlen (KPIs) basiert. Diese Leistungskennzahlen, wie Verfügbarkeit, Auslastung und Reaktionszeit, dienen dazu, die Performance eines Systems objektiv zu bewerten. Das IT-Monitoring spielt dabei eine zentrale Rolle, indem es die Überwachung automatisiert und klare Richtlinien für die Systemüberwachung bereitstellt. Es deckt verschiedene Aspekte der Systemleistung ab: Überwachung der Systemleistung: Erfassung und Analyse von Metriken wie CPU-, Speicher- und Festplattenauslastung, um Engpässe oder Ineffizienzen frühzeitig zu erkennen. Verfügbarkeitsüberwachung: Sicherstellung der Betriebsbereitschaft durch Monitoring von System-Services und -Diensten, die bei Störungen oder Ausfällen automatische Alarme auslösen. Fehler- und Ereignisprotokollierung: Aufzeichnung und Diagnose relevanter Fehler und Ereignisse, die eine Grundlage für detaillierte Analysen bieten. [4] Um IT-Systeme effektiv zu überwachen, ist es sinnvoll, spezifische Schwellenwerte festzulegen, die Alarmmeldungen auslösen, sobald bestimmte Grenzen überschritten werden. Diese Schwellenwerte basieren auf den im System definierten Leistungsindikatoren, wie beispielsweise Prozessaktivitäten, CPU-Auslastung, Speicherauslastung oder Festplattennutzung. Nachdem der normale Systemzustand ermittelt wurde, können die Schwellenwerte entsprechend konfiguriert werden. Der Nutzen dieser Warnmeldungen liegt in der frühzeitigen Identifikation potenzieller Probleme, sodass Administratoren proaktiv Maßnahmen ergreifen können, bevor größere Störungen auftreten. [1]

Umsetzung

Im Rahmen der Bachelorarbeit wurde ein Prozess zur systematischen Aufbereitung der Ausgangsdaten entwickelt. Dieser umfasst zunächst die Analyse und Bereinigung der Daten, bevor sie abschließend für die Validierung aufbereitet werden. Die Transformation findet dabei im Backend statt: Hier werden die Daten aus den Quellsystemen in logische Klassen überführt, um eine effiziente und zuverlässige Validierung zu ermöglichen. Für diese Validierung wurden spezifische Metriken (mit definierten Schwellenwerten) eingeführt, die auf Basis ihrer Vorgaben jeweils einen Statuswert ausgeben. Ein Beispiel ist die Überwachung der Festplattenauslastung:

- Error (rot): ab 99 % Auslastung
- Warning (gelb): zwischen 98 % und 99 % Auslastung
- Normal (grün): unter 98 % Auslastung

Um diese flexible Validierung umzusetzen, wurde im Backend ein Metrik-System entwickelt, das leicht um neue Metriken erweitert werden kann. Die dahinterstehende Logik basiert auf einer sogenannten MetricFactory, die über einen speziellen Backend-Service gesteuert wird. Ergänzend dazu kommen spezialisierte Service-Klassen zum Einsatz, die anhand der definierten Metriken die Daten validieren. Anschließend werden die Validierungsergebnisse über REST-Schnittstellen bereitgestellt. Im Angular-Frontend werden die validierten Daten in Tabellen dargestellt, wobei der Styleguide der 3LiterPPS®-Software verwendet wird. In den Tabellen lassen sich die validierten Systeme und ihre jeweiligen Statuswerte leicht überblicken, wie in der folgenden Abbildung zu sehen ist

Abb. 1: Startseite Reporting-Tool [2]

Darüber hinaus wurde festgelegt, wie die durchgeführten Validierungen dem LF-Consult dabei helfen, zukünftig nachhaltige Lösungen zu entwickeln. Zu diesem Zweck entstand eine Excel-Mockup-Tabelle, in der die Validierungsergebnisse strukturiert aufbereitet und weiter ausgewertet werden. Diese Tabelle gliedert sich in zwei Bereiche . **Automatisch generierte Systemgründe:** hier liefert das System selbst Informationen wie den Namen des betroffenen Systems sowie die fehlgeschlagenen Metriken inklusive der zugehörigen Fehlgründe. **Manuell ergänzte Einträge:** in diesem Bereich dokumentieren Nutzer die ergriffenen Maßnahmen, um ein System wieder in den Status grün zu versetzen. Zudem werden hier die technischen Ursachen des Problems festgehalten. Die strukturierte Erfassung in der Tabelle ermöglicht eine detaillierte Auswertungen, zum Beispiel die Häufigkeit von Problemen: welche Probleme treten besonders oft auf? Die Effektivität von Lösungen: welche Gegenmaßnahmen waren besonders erfolgreich? Und Ursachenanalyse: welche technischen Gründe liegen häufig zugrunde?

Aus diesen Erkenntnissen lassen sich nachhaltige und vorausschauende Lösungen ableiten, um ähnliche Probleme in Zukunft gezielt zu vermeiden.

System Architektur

Als Teil der Bachelorarbeit wurde eine Systemarchitektur entwickelt, die eine automatisierte Validierung von Systemen ermöglicht. Ein Java-Spring-Boot-Backend übernimmt dabei die Strukturierung der eingehenden Daten in verschiedene Klassen und ordnet sie mithilfe eines Thread-Pools anhand eindeutiger System-IDs den jeweiligen Systemen zu. Zusätzlich lädt das Backend die erforderlichen Metrikdaten aus der Datenbank und bereitet sie in einer sogenannten MetricFactory auf. Diese sorgt dafür, dass alle Metriken ein einheitliches Format erhalten und in den passenden Datentyp umgewandelt werden. Auf diese Weise können beispielsweise Datums- oder Prozentwerte, die in der Datenbank zunächst als String vorliegen, korrekt interpretiert werden. Ein Beispiel hierfür ist die Festplattenkapazität, die als Zeichenkette in der Datenbank gespeichert ist und innerhalb der MetricFactory in einen Double-Wert konvertiert wird, um die Auslastungsprozente validieren zu können. Darüber hinaus werden die Metriken in der MetricFactory über ein gemeinsames Interface angelegt. Dieses Interface verfügt über eine Methode, mittels der alle Daten an die Metriken übergeben werden können, um sie anschließend gemeinsam zu validieren. Jede Metrik implementiert dieses Interface und erhält dadurch die Möglichkeit, die ihr zugewiesenen Daten eigenständig zu prüfen. Die Validierung selbst wird durch einen Service angestoßen, der die zuvor eingelesenen Systemdaten an das Interface übergibt. Nach Abschluss der Validierung werden die Ergebnisse in einer sogenannten Result-Klasse gespeichert. Dadurch entsteht eine einheitliche Datenstruktur, die anschließend über verschiedene REST-Endpunkte an das Angular-Frontend übermittelt werden kann.

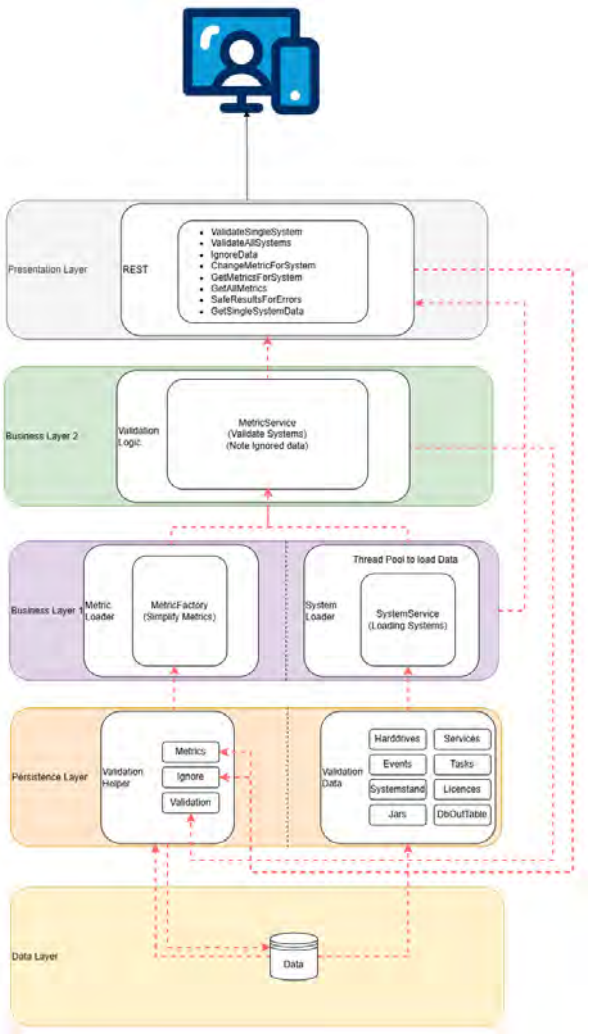


Abb. 2: System-Architektur [2]

Ausblick

Um besser auf die individuellen Eigenschaften der einzelnen Systeme eingehen zu können, ist geplant, das Metrik-System zu erweitern. Dabei sollen die Schwellenwerte der Metriken systembezogen angepasst werden können. Ziel ist es, die bereits vorhandenen Metriken als Standardwerte beizubehalten, wobei diese Standardwerte für jedes System individuell angepasst werden können. Dies bietet den Vorteil, dass Systeme, die aufgrund spezifischer Eigenschaften Daten liefern, die von den Standardwerten abweichen, entsprechend berücksichtigt werden können. Zudem ermöglicht es, gezielt bestimmte Daten oder Systeme zu ignorieren, damit diese nicht mehr in die Validierung einfließen. Dies ist besonders hilfreich, wenn Updates oder Änderungen an den Systemen zu Ergebnissen führen, die ein „Rotes System“ verursachen. Abschließend soll die Auswertung der Excel-Tabellen zukünftig automatisch über das Reporting-Tool erfolgen.

Literatur und Abbildungen

[1] Adam Bertman. Einstieg-in-die-Ueberwachung-mit-Schwellenwerten. https://www.computerweekly.com/de/ratgeber/Einstieg-in-die-Ueberwachung-mit-Schwellenwerten, 2020.

[2] Eigene Darstellung.

[3] Sarah Glöckner. Reporting. https://www.portalsystems.de/wiki/reporting/, 2022.

[4] Markus Vije. IT-Monitoring: Daher ist es so wichtig. https://vije.de/it-monitoring/, 2024.

Vergleich und Implementierung von Konzepten für die Erstellung von Windows Fileless Malware

Enrico Belgiovine **Tobias Heer**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Motivation und Problemstellung

Dateilose Schadsoftware (Fileless Malware) gewinnt in der IT-Sicherheit an Bedeutung, da sie keine Dateien auf der Festplatte ablegt und minimal bis keine Spuren hinterlässt. Sie nutzt legitime Systemprozesse wie PowerShell, um Schadcode im Arbeitsspeicher auszuführen und Spuren zu vermeiden, was die Erkennung erschwert [4]. Die Relevanz zeigt sich in einem Anstieg von 1400% bei Fileless-Angriffen im Jahr 2022 [5]. Angreifer umgehen damit erfolgreich klassische dateibasierte Erkennungsmethoden, da ihre Aktivitäten höchstens im Windows Event Viewer nachverfolgt werden können.

Die zentralen Fragen dieser Arbeit lauten: Wie arbeiten die Komponenten von Fileless Malware zusammen und wie beeinflussen sie sich? Wie lässt sich dieses Zusammenspiel bewerten, um die Effizienz und Schwachstellen der Malware zu analysieren? Hierfür wird untersucht, wie Fileless Malware mithilfe von C# und PowerShell in Windows eingeschleust wird und wie dabei Mechanismen wie das Antimalware Scan Interface (AMSI) und PowerShell-Beschränkungen umgangen werden. Ein Skript wird entwickelt, das die Auswahl und Kombination einzelner Techniken zur Umgehung dieser Mechanismen ermöglicht. Der Fokus liegt auf der Evaluation und Effektivität der Komponenten bei der Umgehung von Sicherheitsmechanismen, insbesondere im Hinblick auf unerkanntes, dateiloses Ausführen zur Datenextraktion nach dem Grundsatz von Fileless Malware.

Aufgrund moderner Sicherheitsmechanismen in Windows 10/11, wie Microsoft Defender und AMSI, sind Techniken wie Verschleierung, Prozessinjektion sowie das Umgehen von Antivirus-Software und virtuellen Maschinen notwendig und werden daher analysiert.

Design

In dieser Arbeit werden Komponenten einer Fileless Malware unter Windows entwickelt, um deren Wechselwirkungen kombiniert als Fileless Malware zu untersuchen, wie in Abb. 1 dargestellt.

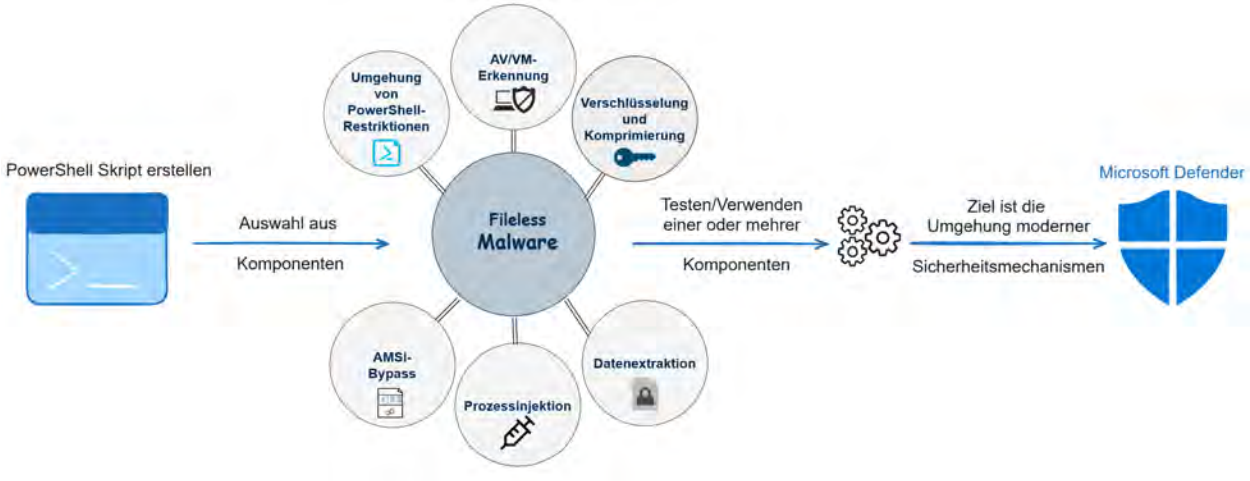


Abb. 1: Fileless Malware Ablauf [2]

Die Angriffstechniken werden in C# entwickelt und über PowerShell mit .NET Unterstützung ausgeführt. Implementiert und gesteuert wird dies durch ein Python-Skript. Folgende Hauptziele werden adressiert:

- Umgehung von PowerShell-Restriktionen: Bypass-Flags wie -ExecutionPolicy Bypass, dynamische Codeausführung über Invoke-Expression werden untersucht, um Restriktionen wie den Constrained Language Mode oder AppLocker zu umgehen.
- AMSI-Bypass: Techniken wie Memory Patching und Fehlererzeugung zur Manipulation des AMSI-Kontexts werden zur Umgehung der Schadcodeüberprüfungen durch AMSI verwendet. Zudem erfolgt der Einsatz von Reflection. Ergänzend werden Obfuskationstechniken wie Base64-Encoding und falls möglich ein PowerShell-Downgrade auf Version 2.0 zur Reduzierung der Erkennungswahrscheinlichkeit genutzt.
- Verschlüsselung und Komprimierung: Der Schadcode wird mit AES-Verschlüsselung und GZIP-Komprimierung verschleiert, um die statische Analyse zu erschweren und die Erkennbarkeit durch signaturbasierte Methoden zu reduzieren. Dies verlängert jedoch die Ausführungszeit durch den Entschlüsselungs- und Entpackungsvorgang.
- AV/VM-Erkennung: Zur Erkennung von Antivirus- und virtuellen Umgebungen werden Systemressourcen, Prozesse und spezifische Dateien untersucht. Zusätzlich wird eine Benutzerinteraktion zur Erkennung automatisierter Analyseumgebungen integriert.
- Prozessinjektion: Self-Injection und Remote Process Injection werden implementiert, wobei letztere für dateilose Angriffe bevorzugt wird. Windows API Funktionen wie *VirtualAllocEx*, *WriteProcessMemory* und *CreateRemoteThread* injizieren Payload in einen laufenden Prozess, um Schadcode auszuführen. Risiken wie Prozessabstürze und erhöhte Erkennbarkeit werden durch gezielte Prozessauswahl minimiert.
- Zugangsdaten und Datenextraktion: Techniken zur Extraktion sensibler Daten werden untersucht, einschließlich des Auslesens des LSASS-Prozesses und der Einsatz von Tools wie *Invoke-Mimikatz.ps1* oder *NativeDump* via Base64-Dekodierung, zur Privilegieneskalation und lateralen Bewegung.

Das im Rahmen dieser Arbeit entwickelte Skript nutzt Parameter und Abfragen, um gezielt Angriffstechniken auszuwählen und zu kombinieren. Bei der Entwicklung werden ein modularer Aufbau und Einschränkungen der C# Funktionalität mit PowerShell berücksichtigt. Ethische Aspekte und ein verantwortungsvoller Einsatz werden einbezogen.

Evaluation

Die Evaluation untersucht die Effektivität der Umgehungstechniken in Windows unter verschiedenen Sicherheitskonfigurationen, insbesondere gegen die Online-Version von Windows Defender und PowerShell-Restriktionen. Zur Überprüfung des AMSI-Bypasses wird AMSITrigger.exe verwendet [7]. Ein Testbeispiel zeigt den Ansatz: Auf einem Windows 10-System mit aktivem Windows Defender injiziert ein PowerShell-Skript Schadsoftware in den LSASS-Prozess, ohne erkannt zu werden. Dabei werden Prozessinjektion, AMSI-Bypass, Obfuskation und dynamische Codeausführung evaluiert. Die Bewertung erfolgt anhand des in Abb. 2 dargestellten Schemas, das inspiriert durch das Bewertungssystem von Arnold et al. [1] für das Thema Fileless Malware angepasst wurde.
Um statistische Signifikanz und Reproduzierbarkeit zu gewährleisten, wird jede Technik mindestens zehnmal einzeln und in Kombination getestet. Es wird eine Zielquote von 80% für hohe Effektivität festgelegt. Außerdem werden qualitative Metriken wie Stabilität (z.B. Prozess-/Systemabstürze), Reproduzierbarkeit und Wechselwirkungen der Techniken bewertet.

Bewertungssystem für Fileless Malware

Kriterium	Technik	Metrik	Beschreibung
Tarnung/Erkennung	AMSI-Bypass und Obfuskation	0 - 100%	Überprüfung der Erkennungsrate mit `AMSITrigger.exe`
Erstellungsaufwand	Verschlüsselung und Komprimierung	0 - 100%	AES- und GZIP-Verschleierung, Erkennungsrate vor/nach Anwendung
Täuschungsfähigkeit	Prozessinjektion und Powershell-Restriktionen	0 - 100%	Erkennungsrate durch Prozessinjektionen und Umgehung von Restriktionen
Unmittelbarkeit	AV/VM-Erkennung (Microsoft Defender/ Hyper-V)	0 - 100%	Tests auf Dateien und Prozesse, die auf virtuelle Maschinen oder Microsoft Defender hinweisen
Exfilteration	Datenextraktion	0 - 100%	Erfolgsrate der unentdeckten Datenextraktion (z.B. Passwörter/Hashes)

Abb. 2: Bewertungssystem für Fileless Malware [2]

Zusätzliche Einblicke bieten die Analyse von Events im Windows Event Viewer, z.B. Event ID 4103 (PowerShell Module Logging) und Event ID 4688 (Audit Process Creation), welche die Erkennungsrate ergänzen. Abschließend werden ethische Aspekte einbezogen, um

das Verständnis zur Verbesserung der Cybersicherheit zu fördern.

Verwandte Arbeiten

Die Forschung zu dateiloser Malware hat in den letzten Jahren zugenommen, da diese Angriffe zunehmend zur Bedrohung werden. Studien beleuchten verschiedene Aspekte und Herausforderungen dieser Technologie.
Liu et al. [3] bieten einen Überblick zur Evolution dateiloser Angriffe sowie Erkennungstechniken und schlagen ein Bedrohungsmodell zur Klassifizierung dieser Methoden vor. Diese Arbeit konzentriert sich im Vergleich dazu auf die Implementierung und Untersuchung spezifischer Angriffstechniken.
Samociuk [6] zeigt, dass die Wirksamkeit von Antivirus-Umgehungstechniken wie Obfuskation und Prozessinjektion häufig mit dem Alter und der Bekanntheit der Tools korreliert. Für seine Tests nutzt er Tools wie Veil-Evasion, Msfvenom und Hyperion, wobei aktuelle Antivirenlösungen die meisten durch diese Tools erzeugten Payloads zuverlässig erkennen und blockieren. Auf diesen Erkenntnissen aufbauend fokussiert sich diese Arbeit auf die Kombination dateiloser Malware-Techniken unter Windows. Samociuks Erkenntnisse werden bei der Implementierung von Obfuskationstechniken einbezogen, ergänzt durch AES-Verschlüsselung und GZIP-Komprimierung zur Tarnung des Schadcodes.
Arnold et al. [1] entwickeln ein Bewertungssystem für PowerShell-Malware, dessen Kriterien teilweise für diese Arbeit übernommen werden. Diese Systematik bildet den konzeptionellen Rahmen für die Implementierung und Evaluierung spezifischer Angriffstechniken, insbesondere zur Tarnung. Kriterien zur Vielseitigkeit und Persistenz werden jedoch ausgeschlossen, da sich die Arbeit auf Windows konzentriert und keine Persistenz getestet wird.
Trotz zunehmender Forschung fehlt eine Analyse der Kombination verschiedener Angriffsmethoden. Im Gegensatz zu früheren Studien integriert diese Arbeit verschiedene Angriffsmethoden in ein Skript und bewertet deren Wirksamkeit unter realen Bedingungen auf aktuellen Windows-Systemen sowie der verantwortungsvolle Umgang mit den Ergebnissen.

Ergebnis

Das Skript stellt die erforderlichen Komponenten zur Analyse und Evaluierung dateiloser Malware unter Windows bereit. Es vereinfacht komplexe dateilose Angriffe im Penetration Testing und unterstützt die effiziente Generierung von Schadsoftware in C# und PowerShell mit hoher Erfolgsrate bei der Umgehung moderner Sicherheitslösungen. Die Evaluation untersucht die Effektivität und Wechselwirkungen dieser Komponenten, um die zentralen Forschungsfragen zu adressieren, die wesentliche Schwachstellen aufdecken und das Bewusstsein für diese Bedrohung stärken. Ethische Gesichtspunkte und der verantwortungsvolle Einsatz des Skripts sind Teil der Analyse zur Verbesserung der Cybersicherheit.

Literatur und Abbildungen

[1] D. Arnold, C. David, and J. Saniie. PowerShell Malware Analysis Using a Novel Malware Rating System. In *Proceedings of the 2022 IEEE International Conference on Electro Information Technology (eIT)*, pages 182–187. IEEE, 2022.

[2] Eigene Darstellung.

[3] Side Liu, Guojun Peng, Haitao Zeng, and Jianming Fu. A survey on the evolution of fileless attacks and detection techniques. *Computers & Security*, 137, 2024.

[4] Steve Mansfield Devine. Fileless attacks: compromising targets without malware. *Network Security*, 4:7–11, 2017.

[5] Aqua Nautilus. Cloud Native Threat Report. https://info.aquasec.com/2023-cloud-native-threat-report, 2023.

[6] Dominik Samociuk. Antivirus Evasion Methods in Modern Operating Systems. *Applied Sciences*, 13, 2023.

[7] Rythm Stick. AMSITrigger: The Hunt for Malicious Strings. https://github.com/RythmStick/AMSITrigger, 11 2020.

Technologische Transformation im IT-EventService: Von der Ist-Analyse zur Entwicklung eines neuen Geschäftsmodells am Beispiel der audius GmbH

Andre Benzinger **Catharina Kriegbaum-Kling**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma audius, Weinstadt

Einleitung

Die COVID-19-Pandemie hat die Veranstaltungsbranche stark beeinflusst und zu einem Rückgang von Präsenzveranstaltungen geführt. Unternehmen mussten schnell auf hybride und Online-Veranstaltungen umstellen, was die Bedeutung einer robusten IT-Infrastruktur unterstrich. [5] In diesem Kontext spielt die digitale Transformation eine entscheidende Rolle, welche starken Einfluss auf die Art und Weise hatte, wie Services und Dienste über ein Netz abgerufen oder bereitgestellt werden. Unternehmen mussten sich den Gegebenheiten anpassen und Ihre Prozesse und Infrastruktur auf neuere Technologien umstellen, wie beispielsweise SaaS-Services von Service-Providern über eine Public-Cloud abzurufen. Diese Veränderung hat den Bedarf nach VPN-Verbindungen am Veranstaltungsstandort zum Konzernnetzwerk erheblich verringert.

Zielsetzung der Arbeit

Die Arbeit zielt darauf ab, die aktuelle Situation und die Auswirkungen der technologischen Transformation auf den IT-EventService zu untersuchen und zu identifizieren, wo Optimierungspotentiale vorhanden sind, um strategische Handlungsempfehlungen abzuleiten, die die Wettbewerbsfähigkeit des Service verbessern könnten. Als Rahmen zur Umsetzung dieses Ziels wurde die Ist-Analyse als passende Methode ausgewählt und an die speziellen Anforderungen des IT-EventService angepasst. In der folgenden Abbildung wird der Ablauf der Ist-Analyse beschrieben.

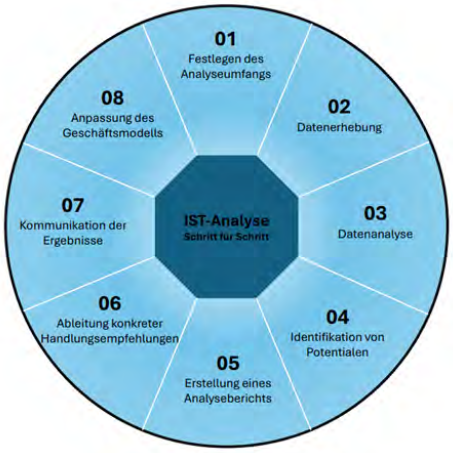


Abb. 1: Ablaufdiagramm Ist-Analyse [2]

IT-EventService

Der IT-EventService stellt sicher, dass IT-Services für Messen, Veranstaltungen und temporäre Anbindungen zur Verfügung stehen. Aufgrund der hohen Außenwirkung einer Veranstaltung erfüllt dieser Service über den Veranstaltungszeitraum sehr hohe Ansprüche bezüglich Qualität, Leistung, Flexibilität und Verfügbarkeit. Mit dem IT-EventService wird sichergestellt, dass Veranstaltungen von Unternehmen in- und außerhalb von Konzernstandorten IT-seitig mit einer hohen Servicequalität im Client- und Netzwerkumfeld ausgestattet werden. Der Clientservice beinhaltet die Bereitstellung und Betreuung der dafür erforderlichen Endgeräte, während der Netzwerkservice die Anbindung an das Internet und die Kopplung mit dem Konzernnetz umfasst.

Technologische Transformation

Die technologische Transformation bezieht sich auf den Prozess der Integration digitaler Technologien in alle Bereiche eines Unternehmens. Dieser Prozess führt zu grund-legenden Veränderungen in der Art und Weise, wie Unternehmen arbeiten und Werte liefern. Technologische Transformation umfasst die Implementierung neuer Technologien wie Cloud-Computing, Künstliche Intelligenz (KI), Internet der Dinge (IoT) und Datenanalytik. Diese Technologien zielen darauf ab, die Effizienz zu steigern, die Flexibilität zu erhöhen und Innovationen voranzutreiben. Obwohl die Begriffe technologische Transformation und Digitalisierung oft synonym verwendet werden, gibt es wichtige Unterschiede zwischen ihnen. Die technologische Transformation ist ein umfassender und strategischer Ansatz, der die gesamte Organisation betrifft und tiefgreifende Veränderungen in der Art und Weise, wie das Unternehmen arbeitet, erfordert. Die Digitalisierung hingegen ist ein taktischer Ansatz, der sich auf die Umwandlung spezifischer Prozesse und Aufgaben konzentriert. [3]

Abb. 2: Prozess technologischer Transformation [1]

Auf den IT-EventService bezogen bedeutet das, dass viele Auftraggeber oder Veranstaltungsorte durch die Technologische Transformation mit einer viel besseren IT-Infrastruktur ausgestattet wurden, was die Relevanz einiger Leistungen des IT-EventService reduziert oder sogar überflüssig macht. In Großen Veranstaltungshallen, wo damals noch kein einziger Router stand, existiert heutzutage oftmals eine komplette WLAN-Infrastruktur mit optimaler Ausleuchtung und Signalstärke.

Analysemethoden

Zur Darstellung der aktuellen Situation und Analyse von Optimierungspotentialen wurden unterschiedliche Analyse- und Datenerhebungsmethoden kombiniert. Durch die Befragung von unterschiedlichen Mitarbeitern des IT-EventService in Form von Experteninterviews mit anschließender Auswertung der Ergebnisse sollte ein Gesamtüberblick hergestellt werden, welcher unterschiedliche Betrachtungspunkte auf den Service berücksichtigt. [4] Über diese sollte die historische Entwicklung, die aktuellen Herausforderungen als auch zukünftige Potentiale durch Meinungen der Experten identifiziert werden und in Form von einer Servicebeschreibung und SWOT-Analyse darstellt werden. Ergänzend zum Experteninterview erfolgte eine Datenerhebung der Aufwandsschätzungen von Projekten aus dem Jahr 2023 mit anschließender Analyse mittels Plan-Ist-Vergleich, Aufwandsanalyse, Wertanalyse und Risikoanalyse. Die Analyse der Datenerhebung hatte das Ziel, Potentiale in den verschiedenen Aktivitäten zu finden, die Bedeutung von Projektkomponenten zu ermitteln, die Planungsqualität zu untersuchen, potenzielle Risiken frühzeitig zu erkennen sowie herauszufinden, welche Tätigkeiten die höchsten Kosten generieren, um schlussendlich Handlungsempfehlungen für die strategische Planung ableiten zu können.

Ausblick

Langfristig gesehen könnte der IT-EventService sein Portfolio erweitern, um innovative Dienstleistungen anzubieten, die auf den neuesten technologischen Trends basieren oder alternative Bedürfnisse decken, wie beispielsweise fortschrittliche Datenanalysen oder Beratungen zur besseren Planung der Netzwerkstruktur und Durchführung von Events bereitzustellen. Zudem wäre eine Prozessanalyse der unterschiedlichen Skill-Level Tätigkeiten empfehlenswert, so lassen sich technische, prozessuale und organisatorische Potentiale ermitteln. Diese Arbeit soll auch die Relevanz von Methoden zur Weiterentwicklung eines Service hervorheben und einen Rahmen für zukünftige Potentialanalysen bieten.

Literatur und Abbildungen

[1] Redaktion AdSimple GmbH. Chancen und Herausforderungen bei der digitalen Transformation. https://www.slashtechnik.de/digitale-transformation-in-unternehmen-was-bringt-sie/, 09 2023.

[2] Eigene Darstellung.

[3] Mark Harwardt. *Technologische Grundlagen der digitalen Transformation*. Springer Gabler Wiesbaden, 2 edition, 2022.

[4] Traute Kaufmann. *Strategiewerkzeuge aus der Praxis: Analyse und Beurteilung der strategischen Ausgangslage*. Springer Gabler, Berlin, Heidelberg, 1 edition, 2021.

[5] Ralf Kunze, Andrea Dessi, et al. Meeting- & EventBarometer 2022/2023. https://www.gcb.de/site/assets/files/78698/management_summary_meeting-_eventbarometer_2022-23.pdf, 05 2023.

Entwicklung und Evaluierung einer Abstandserkennung für ein semi-aktives Exoskelett sowie der Optimierung eines Datenerfassungs-Frameworks

Philip Boehringer **Clemens Klöck**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Fraunhofer IPA, Stuttgart

Motivation

Muskel-Skelett-Erkrankungen (MSE) zählen zu einem führenden Faktor von krankheitsbedingten Arbeitsausfällen in Deutschland [2]. Dabei zählen Arbeiten in Zwangshaltung wie etwa bei Überkopftätigkeiten als Risikofaktor, MSE zu entwickeln [1]. Um diesem entgegenzusteuern, werden Exoskelette als Potenziale gesehen [9], [7]. Vor allem aktive Exoskelette, deren Wirkprinzip auf aktiven Elementen wie etwa Elektromotoren besteht, wirken sich vorteilhaft aus bei schweren, sich wechselnden Lasten [8]. Aktiven Exoskeletten fehlt es derzeit aber noch an geeigneten Ansteuerungsmethoden, um zwischen dem Bewegungswunsch des Nutzers und dem Einfluss der Belastung zu unterscheiden [5], [6]. Ein Ansatz des Fraunhofer IPAs ist hierbei die Verwendung von Informationen der verwendeten Werkzeuge, um die Nutzenden bei deren Tätigkeiten mithilfe eines Exoskeletts bedarfsgerecht zu unterstützen.

Ziel der Arbeit

Das Ziel dieser Bachelorarbeit ist die Entwicklung und Evaluierung eines Systems zur Abstandserkennung, das die Ansteuerung eines semi-aktiven Oberkörper-Exoskeletts verbessert. Das Exoskelett soll bei Überkopftätigkeiten unterstützen. Bisher können nur kinematische Signale und die Aktivierung der Schulter zur Ansteuerung verwendet werden. Um aber ein bedarfsgerechtes Unterstützungsmoment der Schulter zu bestimmen, wird der Abstand (Hebelarm) benötigt. Ein weiterer Schwerpunkt liegt auf der Optimierung eines bestehenden Datenerfassungs-Frameworks (DF), das die erfassten Daten effizienter und genauer verarbeitet und zur Analyse der Experimente mit dem Exoskelett bereitstellt. Die Kombination dieser beiden Ansätze soll zu einer signifikanten Verbesserung der Leistungsfähigkeit und Benutzerfreundlichkeit des Exoskeletts führen. Die Arbeit umfasst sowohl die theoretische Entwicklung als auch die praktische Implementierung und Evaluierung der vorgeschlagenen Lösungen. In Abb. 1 ist die Durchführung der angestrebten Evaluation der Abstandserkennung abgebildet. Dadurch wird ein umfassender Beitrag zur Weiterentwicklung von Exoskelett-Technologien und deren Anwendung in industriellen und medizinischen Bereichen geleistet.

Abb. 1: Simulator zur Evaluation von Bohrübungen mit Exoskelett. Die Werkzeuge und der Simulator senden Daten über MQTT an den Laptop. Auf dem Laptop speichert das Datenerfassungs-Framework die Daten. [4]

Grundlagen Exoskelette

Exoskelette können in drei verschiedene Arten unterteilt werden: passiv, aktiv und semi-aktiv. Passive Exoskelette nutzen keine externen Energiequellen. Stattdessen machen sie sich die Speicherung und Umverteilung von Kräften zunutze. Dafür verwenden sie zum Beispiel Federn, elastische Materialien oder mechanische Konstruktionen, um Bewegungen zu unterstützen. Aktive Exoskelette machen sich externe Energiequellen zunutze. Motoren oder Aktuatoren werden dafür häufig verwendet, um den Bewegungswunsch des Nutzers zu unterstützen. Sensoren und

Mikrocontroller werden verwendet, um diese Unterstützung bedarfsgerecht einzusetzen. Semi-Aktive Exoskelette sind eine Kombination aus den beiden zuvor beschriebenen Arten. Die Unterstützungskraft geschieht durch passive Elemente (Feder). Ein Nachteil ist, dass diese Kraft dauerhaft wirkt und der Nutzer in entgegengesetzter Unterstützungsrichtung mehr Kraft aufwenden muss, um sich bewegen zu können. Deshalb werden aktive Elemente (Motoren) eingesetzt, um diese Unterstützungskraft zu regulieren. Die Regulierung geschieht durch gezieltes Abschalten der passiven Elemente oder durch Variation der Intensität. So können auch stärkere Federn verwendet werden wie bei passiven Exoskeletten, aber die Motoren müssen nicht so groß ausgelegt werden wie bei aktiven Exoskeletten.

Problemstellung und Lösungsansätze

Datenerfassungs-Framework: Für Experimente mit dem Exoskelett wird ein Datenerfassungs-Framework verwendet, das von verschiedenen Werkzeugen, einer Evaluationsplattform (Abb. 1) und dem Exoskelett Daten erfasst und diese speichert, sodass im Anschluss an die Experimente die Daten ausgewertet werden können. Diese Applikation ist essenziell, um Rückschlüsse auf die richtige Funktionsweise des Exoskeletts zu evaluieren. Dafür ist eine zuverlässige Datenübertragung und Speicherung notwendig. Das derzeitige Datenerfassungs-Framework arbeitet mit dem User Datagram Protocol (UDP) und es kann kein Datenverlust der Daten der Werkzeuge und Messeinstrumente erkannt werden. Deshalb soll eine neue Applikation auf der Grundlage des Transmition Control Protocol (TCP) entwickelt werden. Die Entscheidung fiel auf MQTT. MQTT ist eine optimale Wahl, da es auf TCP basiert und zusätzlich Machine-To-Machine-Kommunikation (M2M-Kommunikation) über das Publisher-Subscriber Pattern zur Verfügung stellt. Das ist ein sehr großer Vorteil, da M2M-Kommunikation neue Möglichkeiten für die Abstandserkennung der Werkzeuge zum Exoskelett und der Werkzeugerkennung darstellt. M2M ermöglicht es, dass das Exoskelett mit jedem Werkzeug direkt kommunizieren und Daten austauschen kann. Die Architektur des DF ist eine Multiprozess-Applikation. Wesentliche Komponenten des DF sind ein *MQTTClient*, der kontinuierlich auf Nachrichten der Werkzeuge und der Messinstrumente reagiert und ein *FileHandler*, der für das Speichern der empfangenen Daten zuständig ist. Die Architektur wurde gewählt, damit sich der *MQTTClient* und der *FileHandler* nicht gegenseitig blockieren und eine höhere Empfangsfrequenz sichergestellt werden kann. Die Umstellung auf eine Applikation auf der Basis von TCP ermöglicht es nun, Datenverlust zu verhindern und durch die Multiprozess-Architektur eine zuverlässige und schnelle Datenverarbeitung und -speicherung zu garantieren.

Abstandserkennung: Das derzeitige Problem bei der Unterstützung durch das Exoskelett ist, dass es kein Wissen darüber gibt, wie viel externes Drehmoment das Exoskelett für die Schulter bereitstellen muss. Nur das Gewicht der Werkzeuge ist bekannt, aber nicht die Distanz zum Werkzeug. Das stellt die bedarfsgerechte Unterstützung vor eine große Herausforderung, denn das Drehmoment muss geschätzt werden. Deshalb soll eine Abstandsmessung zwischen der Schulter und dem Werkzeug in der Handfläche eingeführt werden, damit die Distanz zum Werkzeug bestimmt werden kann. Zwei unterschiedliche Konzepte, die gegeneinander bewertet werden sollen, sind Bluetooth Low Energy (BLE) und Ultra Wide Band (UWB). Diese Konzepte messen die Distanz zu einem bestimmten Gegenstück. Das ermöglicht es, dass das ausgesendete Signal nur vom Werkzeug reflektiert oder zurückgesendet wird und nicht von anderen Objekten, die nicht relevant sind. Dies wäre zum Beispiel bei einer Lösung mit Ultraschall der Fall, der zu jedem Objekt, das die Ultraschallwellen reflektiert, eine Distanz misst. Das Konzept, die Distanz mittels UWB zu ermitteln, wird weiter verfolgt, da das Konzept mittels BLE nur eine Schätzung durch die Signalstärke erlaubt. Die Schätzung mit RSSI ist im Vergleich zu der Distanzmessung mit UWB wesentlich unpräziser und damit für diesen Anwendungsfall ungeeignet. Um eine Distanzmessung mit UWB durchführen zu können, wird ein Anchor und ein Tag benötigt. Ein UWB-Sensor kann sowohl als Anchor als auch als Tag fungieren. Der Anchor ist der Fixpunkt und befindet sich an der Schulterkonstruktion des Exoskeletts. Das Tag ist variabel und im Werkzeug verbaut. Die Distanz kann von einem Anchor zu mehreren Tags gemessen werden. Das Prinzip der Distanzmessung mit UWB basiert auf Time of Flight. Dabei wird die Zeit gemessen, in der das ausgesendete Signal vom Anchor bis zum Tag und die Antwort des Tags zurück zum Anchor benötigt. Aus dieser Zeit und der Annahme, dass sich das Signal mit annähernd Lichtgeschwindigkeit fortbewegt, kann die Distanz ermittelt werden. Mithilfe der Distanzmessung soll zusätzlich das derzeitig verwendete Werkzeug erkannt werden. Das ist besonders wichtig, da jedes Werkzeug ein anderes Gewicht besitzt, und somit auch das Drehmoment für die Schulter angepasst berechnet werden muss. UWB benutzt MAC-Adressen, damit der Anchor und jedes Tag eine eindeutige Identifikation besitzen und die Distanzen zugeordnet werden können. Bei der Antwort des Tags an den Anchor wird die MAC-Adresse des Tags mitgesendet und anschließend kann mit einer Lookup-Tabelle das Gewicht des Werkzeugs, in dem sich das Tag befindet, ermittelt werden. Die Abstandsmessung mit der Erkennung, welches Werkzeug sich in der Handfläche befindet, und der

zugehörigen Drehmomentberechnung für die Schulter ist in einem Zustandsautomaten implementiert. Der entwickelte Zustandsautomat ist in Abb. 2 abgebildet.

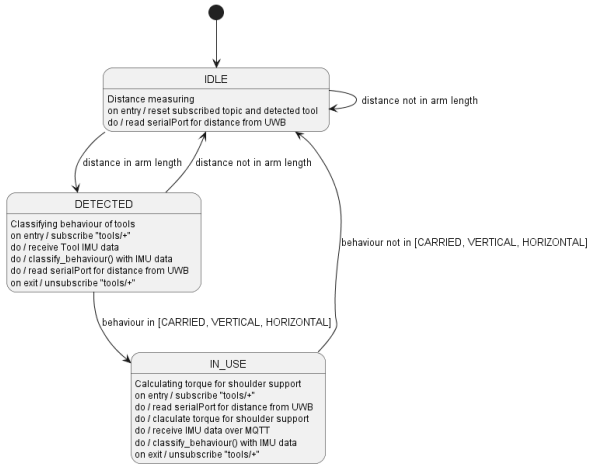


Abb. 2: Zustandsautomat der Abstandsmessung mit Werkzeugerkennung [3]

Evaluation

Die Evaluation für das Datenerfassungs-Framework soll die Datenrate an MQTT-Nachrichten messen, die das Programm verarbeiten und speichern kann. Es sind mindestens 50 Hz gefordert. Zusätzlich soll die maximal mögliche Datenrate ermittelt werden, wenn alle Werkzeuge und Messinstrumente in der maximal möglichen Sendefrequenz senden. Um sowohl die Distanzmessung als auch das DF zu evaluieren, sind Experimente vorgesehen. Die Präzision der Distanzmessung soll evaluiert werden, indem Messungen in vordefinierten Abständen durchgeführt werden. Hierbei wird die vorher bestimmte Distanz über UWB mit der tatsächlichen Distanz verglichen. Anschließend ist ein Experiment geplant, in dem ein Proband das Exoskelett trägt und verschiedene statische Posen mit einem Werkzeug ausführt. Dabei wird die bereits vorhandene Evaluationsplattform (Abb. 1) verwendet. Mit diesen Experimenten soll die Zuverlässigkeit der Distanzmessung sowie die Datenrate des DF evaluiert werden.

Literatur und Abbildungen

[1] J. Barthelme, M. Sauter, C. Müller, and F. Liebers. Association between working in awkward postures, in particular overhead work, and pain in the shoulder region in the context of the 2018 BIBB/BAuA Employment Survey. *BMC Musculoskeletal Disorders, Volume 22, 2021*, 2021.

[2] S. Brennscheidt, A. Siefer, L. Hünefeld, N. Backhaus, and T. Halke. *Arbeitswelt im Wandel*. Bundesanstalt für Arbeitsschutz und Arbeitsmedizin, 2022.

[3] Eigene Darstellung.

[4] Bent Leudesdorff, Lydia Strumpler, Thomas Dobosz, Christophe Maufroy, Urs Schneider, and Thomas Bauernhansl. Sensor System for Real-time Classification of Manual Construction Tasks with Power Tools for Exoskeleton Control. *2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (In Press)*, 2024.

[5] N. Lotti et al. Myoelectric or Force Control? A Comparative Study on a Soft Arm Exosuit. *IEEE Transactions on Robotics (Volume: 38, Issue: 3, June 2022)*, 2022.

[6] C. Luna et al. Admittance-based Upper Limb Robotic Active and Active-Assistive Movements. *International Journal of Advanced Robotic Systems*, 2015.

[7] P. Maurice et al. Objective and Subjective Effects of a Passive Exoskeleton on Overhead Work. *IEEE Transactions on Neural Systems and Rehabilitation Engineering: A Publication of the IEEE Engineering in Medicine and Biology Society*, 2020.

[8] B. Otten et al. Evaluation of a Novel Active Exoskeleton for Tasks at or Above Head Level. *IEEE Robotics and Automation Letters*, 2018.

[9] J. Theurel and K. Desbrosses. Occupational Exoskeletons: Overview of Their Benefits and Limitations in Preventing Work Related Musculoskeletal Disorders. *IISE Transactions on Occupational Ergonomics and Human Factors*, 2019.

Record and Replay of IPC communication in AUTOSAR Adaptive

Wolfgang Bradfisch **Mirko Sonntag**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Vector Informatik GmbH, Regensburg

Introduction

The increasing complexity of electrical and electronic (E/E) systems in vehicles, driven by the integration of more and more sensors, is accompanied by a corresponding increase in the amount of software for infotainment, advanced driver assistance systems (ADAS) and sensor fusion. As a result, the increasing complexity of modern vehicles requires the use of high performance computing (HPC). The prevailing approach in automotive design is to consolidate software operations by running multiple applications throughout the vehicle on a single electronic control unit (ECU). This centralization is enabled by inter-process communication (IPC), which reduces the total number of ECUs required for the relevant software functions. On these HPC ECUs, it is common to have multiple applications that need to share information via IPC communication [2].

In order to standardize software development in the automotive context, the AUTOSAR (AUTomotive Open System ARchitecture) standard was developed. The primary objectives of AUTOSAR are to enhance scalability for a range of automotive applications, facilitate the transferability of software components across diverse systems and platforms, and foster collaboration among automotive manufacturers, suppliers, and developers. This standard enables various companies to contribute different parts of the software efficiently.

AUTOSAR consists of two principal platforms. The initial platform is AUTOSAR Classic, which features a layered architecture. This architecture includes the Basic Software (BSW) layer, which comprises essential drivers and services that are necessary for the vehicle's operation. The Software Components (SWC) layer contains the specific applications that have been tailored for the vehicle's functionality. Finally, the Real-Time Environment layer manages communication between the Basic Software and the Software Components, ensuring seamless operation and coordination within the vehicle's system [5].

The second platform is the AUTOSAR Adaptive Platform, which serves as middleware on a POSIX-compliant system. The key characteristics of the Adaptive Platform include its reliance on a POSIX-compliant OS, which provides a robust foundation for complex applications. The Adaptive Runtime for AUTOSAR (ARA) handles communication between different software components, supporting a service oriented architecture and acting as middleware [3].

It is of paramount importance to verify that applications function as intended and maintain vehicle safety, particularly in the case of ADAS algorithms. The testing of these applications necessitates the utilisation of real-world data, which can be achieved through the playback of recorded data to the ECU. In instances where applications are distributed across disparate ECUs, the recording and replay of bus communications is an adequate methode. However, in scenarios where multiple applications are resident on a single ECU, and the objective is to test a singular application, it becomes imperative to record and replay the IPC communication specific to that application. The objective of this work is to facilitate the recording and replaying of IPC communications.

IPC communication in AUTOSAR Adaptive

In the context of AUTOSAR, communication is service-oriented, distinguishing between a service provider (server) and a service consumer (client). Unlike traditional signal-based systems, which require predefined knowledge of the signals to be communicated, service-oriented communication allows for the request of services as needed. This flexibility enables the incorporation of new services into existing vehicle systems over time.

To access a service, the client must subscribe to the server. This communication uses the Scalable Service-Oriented Middleware over IP (SOME/IP) protocol.

Interaction within this framework is facilitated through the use of a skeleton and a proxy. Crucially, SOME/IP communications include a header containing the message ID, message length, message type, and other pertinent information.
Service discovery (SD) is employed to locate specific services. Once a service is identified, AUTOSAR utilizes fields, methods, and events for IPC communication. Events disseminate information about system changes or specific occurrences, fields represent the data elements exchanged between software components within the AUTOSAR Adaptive framework, and methods illustrate the functions or operations that can be invoked, enabling component interaction and various actions [4].
To test applications that communicate via IPC, it is essential to have the capability to record and replay the communication with sample data. This capability allows for thorough testing, similar to the methods used when testing communication between ECUs over a bus system. The next chapter will explore this subject in greater detail.

Record and Replay

The capacity to record and replay data is of paramount importance for the testing and debugging of complex automotive software systems. In order to achieve this, it is necessary to record the data in its most pristine form, which entails that it should not be serialized within the skeleton. By avoiding data serialization, it is possible to log and replay raw data, thereby simplifying the process and allowing for the use of synthetic data without the need for serialization. The approach involves recording data at the point of sending and replaying it as required by the application, thereby bypassing lower-level communication issues. This ensures that the application receives the intended data, while reducing the data size by logging only the necessary values, rather than the serialized data.
This method allows for the replay of recorded data from any location, provided that the data and a timestamp are available, along with the target server. It is important to note, however, that the precise timestamp of message reception is not captured, which could potentially be a limitation in certain scenarios. Nevertheless, the method offers an optimal compromise between comprehensive testing and minimal lower-level interference, thus obviating the necessity for stack involvement in the replay pipeline. In AUTOSAR Adaptive, modifications are required to the SD component since there is no sending server during replay. Additionally, changes must be made to the proxy's send function to facilitate recording and to the skeleton to enable replay functionality. This can be seen in figure 1 [4].
These modifications guarantee the accurate recording and replay of data, thus providing an effective means for testing and debugging applications within the AUTOSAR Adaptive framework. This method offers a more efficient approach, reduces the need for extensive data storage, and ensures that the replayed data closely resembles the original communication, thereby enhancing the reliability of testing scenarios. The ultimate objective is to develop a concept and prototype for replaying IPC communication in AUTOSAR Adaptive, enabling the testing of complex applications with real-world replay data.

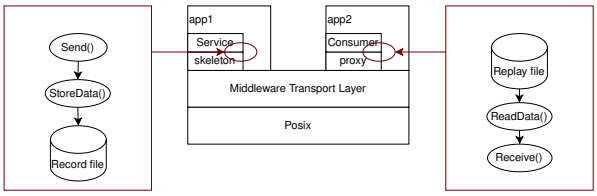


Abb. 1: Record and Replay of IPC communication in AUTOSAR Adaptive [1]

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Simon Fürst and Markus Bechter. AUTOSAR for connected and autonomous vehicles: The AUTOSAR adaptive platform. In *2016 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshop (DSN-W)*, pages 215–217. IEEE, 2016.

[3] AUTOSAR Organization. *Explanation of Adaptive Platform*. AUTOSAR Adaptive, 2024.

[4] AUTOSAR Organization. *Explanation of ara::com API*. AUTOSAR Adaptive, 2024.

[5] AUTOSAR Organization. *Software Component Template*. AUTOSAR Classic, 2024.

Machbarkeitsstudie zur mobilen Signalverarbeitung und Datenanalyse der DC-Ladekommunikation gemäß ISO 15118

Jonas Burger **Dennis Grewe**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung und Problemstellung

Die fortschreitende Elektromobilität erfordert eine effiziente, interoperable und sichere Kommunikation zwischen batterieelektrischen Fahrzeugen (BEV) und Ladestationen. Der ISO-15118-Standard stellt hierfür eine zentrale Kommunikationsschnittstelle dar, die unter anderem die Implementierung von Vehicle-to-Grid (V2G)-Funktionen ermöglicht. Der Standard unterstützt Funktionen wie Authentifizierung, Aushandlung von Ladeparametern (z. B. Leistung, Dauer, Kosten) und sicheres Zahlungsmanagement. ISO 15118 ermöglicht Funktionen wie "Plug and Charge", bei der der Ladevorgang automatisch startet, sowie die sichere Übertragung von Authentifizierungs- und Abrechnungsdaten. Zudem unterstützt ISO 15118 die intelligente Laststeuerung und bidirektionale Kommunikation zur Förderung der Netzstabilität. Damit verbessert der Standard die Effizienz, Sicherheit und Benutzerfreundlichkeit von DC-Ladesystemen und treibt die Weiterentwicklung der Ladeinfrastruktur voran. [3]

Die auf Powerline Communication (PLC) basierende Control-Pilot-Leitung (CP) zwischen Elektrofahrzeug (EV) und Ladestation ist eine Punkt-zu-Punkt-Verbindung und kein Bussystem. Dies erschwert das Abgreifen der Kommunikationssignale, da keine standardisierte Schnittstelle vorhanden ist. Zum Mithören der auf der CP-Leitung übertragenen Daten sind spezialisierte Hardwarelösungen wie der CCS Listener von Vector erforderlich, die die Signale in auswertbare Formate umwandeln. [2] Solche Lösungen sind erfahrungsgemäß komplex, groß, schwer und teuer, was ihren Einsatz im mobilen Feldbetrieb unattraktiv macht.

Konzept & Herangehensweise

Ziel der Arbeit ist die Entwicklung eines kompakten, kosteneffizienten und mobilen Testers, der es ermöglicht, die High-Level-Kommunikation nach ISO 15118 im Feld zu analysieren. Dafür werden zwei Hauptansätze betrachtet:

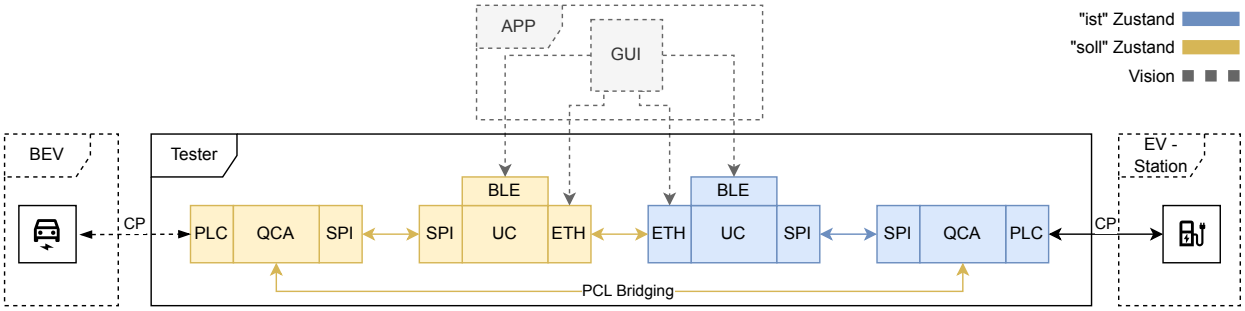


Abb. 1: Projektplan: Signalverarbeitung und Prozessierung [1]

- **Direkte Signalverarbeitung über Mikrocontroller (µC)**

  Hierbei werden alle Signale zwischen BEV und Ladestation über einen Qualcomm Atheros (QCA)-Chip für PLC an den Mikrocontroller geleitet, verarbeitet und wieder zurückgewandelt (siehe Abbildung 1). Dies ermöglicht die Auswertung von Timings und Metadaten. Dadurch sind unvermeidliche Verzögerungen (Latenzen)

zu erwarten, die gemessen und bewertet werden müssen. Ziel ist es, die Kommunikation durchzuleiten und gleichzeitig relevante Informationen auszulesen.

- **Datenbrückung und Filterung durch QCA-Chips (Tunnelung)**

  Im zweiten Ansatz wird angestrebt, bestimmte Komponenten zu überbrücken, um ausschließlich die relevanten Datenströme auszuwerten. Dadurch sollen die Einflüsse auf die Kommunikation reduziert und idealerweise geringere Latenzen erreicht werden. Dazu soll der Mikroprozessor überbrückt und das Signal direkt an den QCA-Chip weitergeleitet werden. Durch diese Form der Tunnelung wird die Kommunikation möglichst unverfälscht übertragen, während Messwerte und Metadaten dennoch erfasst werden können.

Zur praxisnahen Evaluation der entwickelten Ansätze zur Analyse der ISO-15118-Kommunikation ist ein realistisches und kontrolliertes Testumfeld erforderlich. Dieser Aufbau ermöglicht die Simulation der Interaktion zwischen EV und Ladestation unter verschiedenen Bedingungen. So können reale Kommunikationsszenarien nachgebildet und die Effektivität der Signalverarbeitungs- sowie Datenbrückungsmethoden gezielt bewertet werden.

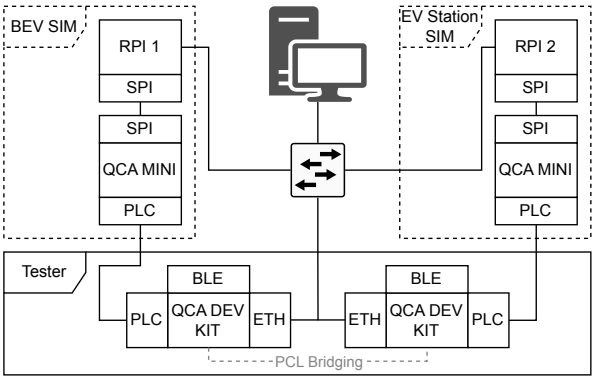


Abb. 2: Testaufbau: Hardwareaufbau des Testers und Simulation von EV und Ladestation [1]

Gemäß Abbildung 2 werden zur Simulation von BEV und Ladestation zwei Raspberry Pis, jeweils ausgestattet mit einem SPI-zu-PLC-Modul (PLC Stamp Mini 2 mit QCA7005 Chip), verwendet. Die Geräte sind über eine PLC-Leitung miteinander verbunden. Als Tester werden, zur Umwandlung der PLC-Kommunikation in Ethernet, zwei Devolo dLAN Green PHY Eval Boards II mit QCA7000-Chips in die PLC-Leitung eingehängt. Zur Analyse werden Messpunkte wie Switches, Router oder Netzwerk-TAPs eingesetzt, die Kommunikationsdaten abgreifen. Diese werden mit Tools wie *Wireshark* oder *tcpdump* aufgezeichnet und ausgewertet. Realistische Umgebungsbedingungen wie Kabellängen und Signalstörungen werden berücksichtigt, um den Feldbetrieb möglichst genau nachzubilden. Eine direkte PLC-Verbindung zwischen den Eval Boards kann zur Minimierung von Latenzen ebenfalls genutzt werden. Ziel des Testaufbaus ist es, die Effektivität und Präzision des Testers zu überprüfen. Dabei soll gezeigt werden, dass ISO-15118-Kommunikationsflüsse praxisnah erfasst, ausgewertet und unverfälschte Metadaten bereitgestellt werden, um die Eignung des Systems für den mobilen Feldeinsatz zu belegen.

Evaluation

Für die Evaluation wird ein realistisches Testszenario aufgebaut. Hierfür wird der Ladevorgang initiiert, authentifiziert und gesteuert, so wie es im realen Betrieb erwartet wird. Es werden ISO-15118-konforme Nachrichten ausgetauscht, um beispielsweise den Ladezustand oder Parametrierungen der Ladung (Leistung, Dauer) zu verhandeln. Hierbei stehen folgende Metriken im Vordergrund:

- Latenzzeiten: Der ISO-15118-Standard definiert spezifische Antwortzeiten, die während der Kommunikation zwischen BEV und Ladestation einzuhalten sind. Daher ist die Messung der Latenzzeiten ein zentrales Kriterium. Sie stellt sicher, dass der Tester die Kommunikationsflüsse ohne signifikante Verzögerungen verarbeiten kann und die angewandten Ansätze den Standardanforderungen entsprechen.
- Integrität der Daten: Lassen sich alle relevanten ISO 15118-Nachrichten vollständig erfassen?
- Mobilität und Handhabbarkeit: Lässt sich der Tester leicht in unterschiedlichen Umgebungen einsetzen?
- Kosten- und Platzersparnis: Welche Vorteile bietet das System gegenüber herkömmlichen Laborsystemen?

Im Rahmen der Machbarkeitsstudie wird untersucht, inwieweit die beiden Ansätze hinsichtlich ihrer Effizienz, Genauigkeit, Stabilität und Praktikabilität realisierbar sind. Ziel ist es, den Ansatz zu identifizieren, der sich am besten für den mobilen Feldeinsatz eignet.

Zusammenfassung und Ausblick

In dieser Bachelorarbeit werden zwei Ansätze zur mobilen Analyse der ISO 15118-High-Level-Kommunikation untersucht: die Direkte Signalverarbeitung über Mikrocontroller (µC) und die Datenbrückung und Filterung durch QCA-Chips (Tunnelung). Die beiden Ansätze

bestehen darin, einerseits die Signale direkt über einen Mikrocontroller zu verarbeiten, um Timings und Metadaten auszuwerten, und andererseits die Datenbrückung und Filterung mittels QCA-Chips durchzuführen, um die Kommunikationslatenzen zu minimieren und eine unverfälschte Übertragung zu gewährleisten. Beide Vorgehensweisen sollen in einem realistischen Testszenario evaluiert und hinsichtlich Latenzen, Effizienz und Benutzbarkeit bewertet werden.
Die in dieser Arbeit gewonnenen Ergebnisse können genutzt werden, um den Tester weiter zu verbessern und zusätzliche Analysefunktionen zu integrieren. Dies ermöglicht eine noch effektivere und benutzerfreundlichere Anwendung im mobilen Feldeinsatz.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Vector Informatik GmbH. VH5110A CCS-Ladekommunikation einfach analysieren. https://cdn.vector.com/cms/content/products/VH5110/Docs/VH5110A_FactSheet_DE.pdf, 06 2024.

[3] Nationale Plattform Zukunft der Mobilität. ROADMAP ZUR IMPLEMENTIERUNG DER ISO 15118. https://www.plattform-zukunft-mobilitaet.de/wp-content/uploads/2020/12/NPM_AG5_AG6_2020_Q4_ISO15518.pdf, 12 2020.

Entwicklung eines Bewertungsinstruments zur objektiven Bewertung von IT-Projekten: Identifikation und Gewichtung spezifischer Kriterien

Lorenzo Carrabba **Thomas Rodach**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma TTS Tooltechnic Systems AG & Co. KG, Wendlingen

Einleitung

Die Digitalisierung hat einen signifikanten Einfluss auf Unternehmen und erhöht den Druck zur erfolgreichen Umsetzung von IT-Projekten. Häufig scheitern solche Projekte jedoch an der Einhaltung von Zeit-, Kosten- und Qualitätszielen. Ein strukturiertes Bewertungsinstrument kann dazu beitragen, Erfolgsaussichten zu erhöhen, Risiken frühzeitig zu erkennen und Entscheidungen besser abzusichern. Abbildung 1 zeigt den wachsenden Wettbewerbsdruck und die Bedeutung technologischer Fortschritte in verschiedenen Branchen. [5]

Abb. 1: Digitalisierung verschärft den Wettbewerb [5]

Zielsetzung

Das Ziel der vorliegenden Arbeit ist es, spezifische Bewertungskriterien für IT-Projekte zu identifizieren und zu gewichten, um ein praxisorientiertes Bewertungsinstrument zu entwickeln. Dieses Instrument soll eine objektive Bewertung von IT-Projekten ermöglichen und dabei sowohl auf wissenschaftliche Grundlagen als auch auf praktische Anforderungen eingehen. Der Fokus liegt insbesondere auf der Identifikation und Gewichtung der Kriterien, während die praktische Anwendbarkeit des Instruments durch eine prototypische Implementierung und Validierung untersucht wird.

Theoretische Grundlagen

IT-Projekte und ihre Besonderheiten

IT-Projekte zeichnen sich durch hohe Dynamik, Komplexität und Unsicherheiten aus. Typische Herausforderungen umfassen die Schätzung des Aufwands, die Überwindung von Kommunikationsbarrieren zwischen IT- und Fachabteilungen sowie technische Abhängigkeiten. Diese Aspekte verdeutlichen die Notwendigkeit eines durchdachten Bewertungsansatzes [4].

Bedeutung der Bewertung

Die Bewertung von IT-Projekten umfasst die Überwachung und Steuerung während des gesamten Projektverlaufs. Sie hilft, Zielabweichungen zu identifizieren und Gegenmaßnahmen einzuleiten. Kriterien wie strategische Relevanz, technische Machbarkeit, Lebenszykluskosten und Risikopotenzial spielen hierbei eine entscheidende Rolle [4].

Existierende Modelle

Zur Entwicklung des Bewertungsinstruments wurden wissenschaftliche Theorien wie Total Cost of Ownership (TCO) und Return on Investment (ROI) sowie praxisorientierte Methoden wie die Nutzwertanalyse (NWA) und die Balanced Scorecard (BSC) analysiert. Diese Modelle bieten wertvolle Ansätze für die Strukturierung der Bewertung [6] [1].

Methodische Vorgehensweise

Die Arbeit kombiniert theoretische Grundlagen und empirische Forschung. Der Prozess gliedert sich in drei Phasen: 1. Identifikation relevanter Bewertungskriterien: Mithilfe einer Literaturrecherche und der Analyse bestehender Modelle wurden vier Hauptkategorien definiert: strategische Relevanz, technische Machbarkeit,

Lebenszykluskosten und Risikopotenzial. Diese Kategorien umfassen 24 Kriterien. 2. Validierung durch eine Unternehmensbefragung: Eine Survey innerhalb der IT-Abteilung validierte die Relevanz und Gewichtung der Kriterien. Die Teilnehmer bewerteten diese auf einer vierstufigen Likert-Skala [3]. 3. Erstellung eines Bewertungsinstruments: Basierend auf den validierten Kriterien wurde ein Scoring-System entwickelt, das Projekte mit einer maximalen Punktzahl bewertet.

Gewichtung und Priorisierung

Die Gewichtung der Kategorien wurde durch die Umfrage bestimmt. So erhielt z. B. eine Kategorie die Gewichtung von 30%, während eine andere Kategorie mit 20 % etwas niedriger bewertet wurde. Die Gesamtpunktzahl einer Kategorie wird dann ins Verhältnis zur maximalen erreichbaren Punktzahl gesetzt, um die prozentuale Gewichtung zu ermitteln. Innerhalb der Kategorien wurden weniger relevante Kriterien ausgeschlossen, um das Instrument kompakt und anwendbar zu gestalten [3]. Das Punktesystem zur Gewichtung der Kategorien ist wie folgt strukturiert:

Unwichtig	0 Punkte
Weniger wichtig	1 Punkt
Wichtig	2 Punkte
Sehr wichtig	3 Punkte

Abb. 2: Punktesystem Gewichtung der Kategorien [2]

Ausblick

Das Bewertungsinstrument bietet eine strukturierte Grundlage für die Auswahl und Steuerung von IT-Projekten. Es wurde so gestaltet, dass es flexibel auf unterschiedliche Projekttypen anwendbar ist und Unternehmen bei strategischen Entscheidungen unterstützt. Zukünftige Arbeiten könnten die Anwendung auf andere Branchen ausweiten oder KI-gestützte Analysetools integrieren, um die Bewertung weiter zu automatisieren und zu präzisieren.

Literatur und Abbildungen

[1] Walther Busse von Colbe and Frank Witte. *Investitionstheorie und Investitionsrechnung*. Springer Berlin, Heidelberg, 5 edition, 2018.

[2] Eigene Darstellung.

[3] Mathias Jesussek. likert-skala. https://datatab.de/tutorial/likert-skala, 2024.

[4] Jens Kaufmann and Wilhelm Müldur. *Grundkurs Wirtschaftsinformatik - Eine kompakte und praxisorientierte Einführung*. Springer Vieweg Wiesbaden, 10 edition, 2023.

[5] Christopher Meinecke and Andreas Streim. Unternehmen-wollen-Digitalisierung-vorantreiben. https://www.bitkom.org/Presse/Presseinformation/Unternehmen-wollen-Digitalisierung-vorantreiben, 2024.

[6] Alexis Savkin. bscdesigner: bsc-vorteile-und-nachteile. https://bscdesigner.com/de/bsc-vorteile-und-nachteile.htm, 2020.

Diagnose von Komponenten der elektrischen Automatisierung durch mobile und cloud-basierte Anwendungen als Erweiterung bestehender PC-Anwendungen unter Berücksichtigung von Cyber-Security Anforderungen

Alexander Dietrich **Jörg Nitzsche**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Festo SE & Co. KG, Esslingen a. N.

Motivation

"There are only two types of companies: those that have been hacked and those that will be."[6]. Das Zitat des ehemaligen FBI-Leiters Robert Mueller unterstreicht die Wichtigkeit der Security in der Realisierung von Softwareanwendungen. Es zeigt auf, dass viele Unternehmen die Gefahr von Cyberangriffen unterschätzen. Die europäische Kommission hat dies erkannt und den Cyber Resilience Act im November 2024 veröffentlicht [3]. Dieser bewirkt, dass Unternehmen die Cyber-Security mit erhöhter Dringlichkeit und strategischer Priorität betrachten. Aus diesem Grund wird dieses Thema bei Festo SE & Co. KG (kurz: Festo), bei der diese Bachelorarbeit verfasst wird, mit besonderer Wichtigkeit behandelt. Insbesondere der Sicherheit in den Automatisierungssystemen des Unternehmens kommt aus diesem Grund eine große Bedeutung zu. In Kombination mit der Entwicklung zu mobilen, cloudbasierten Anwendungen stellt sich die Frage: „Ist die Security in der elektrischen Automatisierung im bestehenden System ausreichend und auf welche essenziellen Aspekte muss bei der Umsetzung und Implementierung mobiler, cloudbasierter Anwendung geachtet werden, um ein hohes Maß an Cyber-Security und Resilienz zu gewährleisten?“. Festo entwickelt und verkauft Automatisierungsgeräte weltweit. Diese Geräte müssen in Betrieb genommen und gewartet werden. Um dies den Kunden so einfach wie möglich zu gestalten, hat Festo eine eigene Software namens Festo Automation Suite (kurz: FAS) entwickelt. Kunden können hiermit, u.a. Geräte in Betrieb nehmen, die Parametrierung durchführen und Diagnosedaten auslesen.

Einleitung

In dieser Arbeit wird das Auslesen von Diagnosedaten in Bezug auf die obige Fragestellung untersucht. Dazu wird mithilfe einer Threat and Risk Analysis (kurz: TARA) das Security Level der Anwendung in diesem spezifischen Fall betrachtet. Darüber hinaus wird diese zu betrachtende Funktionalität der FAS durch eine cloudbasierte App erweitert, welche die Diagnosedaten in einem Dashboard anzeigt und – ebenfalls in der mobilen Anwendung – eine Schritt-für-Schritt-Anleitung zur Fehlerbehebung bietet.

Grundlagen

Die FAS ist eine Software von Festo zur Einrichtung, Parametrierung, Steuerung und Diagnose von hauseigenen Automatisierungskomponenten und basiert auf .NET. Die Software ermöglicht es, Produktfamilien als Plugin zu installieren, und schlägt die optimale Designstruktur der Automatisierungskomponenten vor [5]. Ein Plugin ist dabei eine gerätespezifische Komponente. Es wird mitunter zur Inbetriebnahme, Parametrierung, Kommunikation, Diagnose und zum Speichern, sowie zum Laden von Projekten benutzt. In dieser Arbeit wird ein CMMT-ST, ein Servoantriebsregler von Festo, verwendet. Er kann verwendet werden, um Schrittmotoren oder bürstenlose Gleichstrommotoren zu betreiben. Der CMMT-ST ist sehr gut für Aufgaben geeignet, die einen geringen Leistungsbedarf haben. Darüber hinaus wird er für die genaue Ansteuerung bestimmter Bewegungspunkte verwendet [7]. Für die Parametrierung, sowie den Datenaustausch und die Kommunikation mit dem CMMT-ST wird ein proprietäres Protokoll, das ENGP, verwendet. Dieses Protokoll ist auf dem ISO/OSI-Schichtenmodell auf der

fünften Ebene einzuordnen. ENGP kann ausschließlich synchron kommunizieren. Technologisch setzt es das Vorhandensein der Kommunikationsschicht voraus. Diese muss verbindungsorientiert und transaktionssicher sein, wie es zum Beispiel TCP/IP erfüllt [4].

TARA

Die TARA wird von Security-Experten durchgeführt und hat das Ziel, Sicherheitsrisiken in Projekten oder Organisationen frühzeitig zu erkennen. Hierfür werden Anwendungen basierend auf ihren Datenflüssen analysiert, ausgewertet und in verschiedene Gefahrenstufen eingestuft. Für diverse identifizierte Bedrohungen werden deren Wahrscheinlichkeiten, basierend auf der hierfür benötigten Motivation des Angreifers, sowie die potenzielle Schwere des Angriffs eingeschätzt. In dieser Arbeit wird die TARA mithilfe eines IEC62443 basierten, firmeninternen Ansatzes durgeführt. Dazu wurde ein Datenflussdiagramm des bestehenden Systems entworfen (s. Abb. 1). Auf der rechten Seite ist der Datenfluss innerhalb des CMMT-ST Plugins zu sehen. Das Plugin wird von der FAS gehostet. Die Drives ENGP Communication-Schnittstelle ist Teil des Plugins. Sie ist für die Kommunikation zwischen dem Plugin und dem physischen Gerät (hier: CMMT-ST) zuständig. Die Schnittstelle fragt über das ENGP Diagnosedaten des CMMT-ST ab. Nach Bestätigung der Verbindung werden die Diagnosedaten empfangen und geparst. Damit können sie von der Diagnosis UI abgefragt und angezeigt werden.

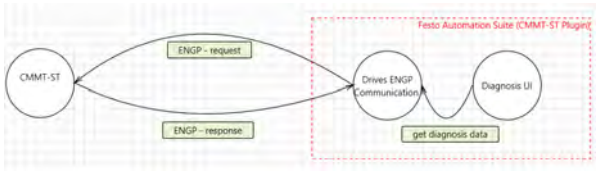


Abb. 1: Datenflussdiagramm [2]

Als nächsten Schritt werden mithilfe des Datenflussdiagramm, Threats identifiziert und nach den STRIDE-Kategorien klassifiziert [1]. Anschließend werden diese basierend auf internen Klassifizierungsvorgaben einer Angriffswahrscheinlichkeitsstufe und einer Schweregradstufe zugeordnet. Zusammen ergeben diese Stufen ein Gesamtrisiko eines Threats. Besonders hervorzuheben ist dabei ein Threat, dessen Risiko als nicht akzeptabel eingestuft ist: Er beschreibt das Problem der fehlenden Handlungsanweisung bei Diagnosedaten. Konkret könnte ein Gerät unabsichtlich beschädigt werden, wenn ein unerfahrener Mitarbeiter vorschnell handelt, indem er beispielsweise falsche Firmware-Updates aufspielt oder andere Wartungsaufgaben inkorrekt durchführt. Er wird nach den STRIDE-Kategorien als ein Denial-of-Service behandelt. Ziel dieser Arbeit ist es, eine cloudbasierte App zu entwickeln um dieses Risiko zu mitigieren.

Konzept und Realisierung

Um direkt über die App in Echtzeit mit Geräten zu kommunizieren, wurde ein MVVM-Entwurfsmuster für die Architektur gewählt. .NET MAUI wurde als Framework gewählt, um weiterhin in der .NET-Umgebung zu bleiben. Das ermöglicht die Integration der bereits vorhandenen Interfaces, um mit dem CMMT-ST zu kommunizieren. Die App ermöglicht es Benutzern, Diagnosedaten von einem Festo-Gerät abzufragen. Diese werden dann als unabhängige Elemente angezeigt. Durch das Drücken einer Meldung gelangt der Nutzer auf eine detaillierte Darstellung dieser. Hier werden automatisiert Fehlerbehebungsmaßnahmen aus einer cloudbasierten PostgreSQL-Datenbank dargestellt (s. Abb. 2).

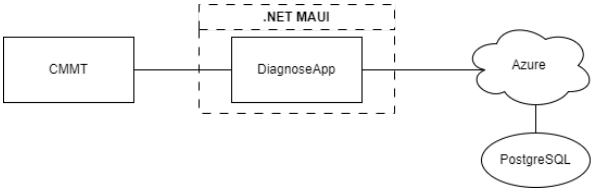


Abb. 2: Softwarearchitektur [2]

Diese geben dem Benutzer eine Schritt-für-Schritt Anweisung, um das Problem zu beheben.
Die App besteht dabei aus drei Seiten. Diese sind in Abb. 3 ersichtlich:

- Links ist die Startseite, welche den Verbindungsaufbau übernimmt, zu sehen.
- In der Mitte ist die Diagnose-Seite mit den jeweiligen Diagnosedaten ersichtlich.
- Rechts ist die Solution-Seite mit der detaillierteren Ansicht der Diagnosemeldung und die Fehlerbehebungsmaßnahmen zu einem spezifischen Problem dargestellt.

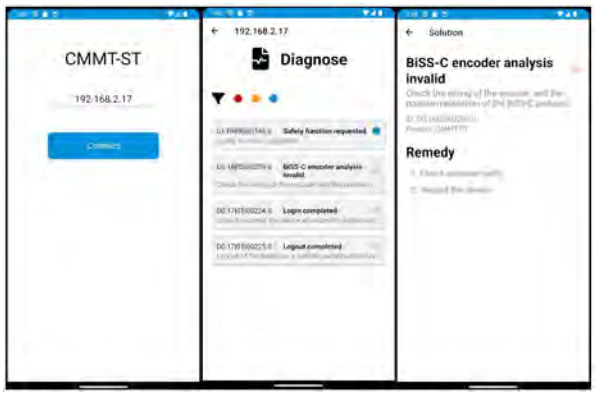


Abb. 3: Benutzeroberfläche der App [2]

Ausblick

Als nächster Schritt in dieser Arbeit wird eine TARA für die App durchgeführt, um diese mit der ersten TARA des bestehenden Systems zu vergleichen. Ziel ist es, aus dieser Analyse eine abschließende Antwort auf die Fragestellung dieser Arbeit abzuleiten.

Literatur und Abbildungen

[1] Microsoft Corporation. The STRIDE Threat Model. https://learn.microsoft.com/en-us/previous-versions/commerce-server/ee823878(v=cs.20)?redirectedfrom=MSDN, 12 2009.

[2] Eigene Darstellung.

[3] Council of the European Union European Parliament. Cyber Resilience Act. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R2847, 11 2024.

[4] Armin Hartmann. ENGP - Codebeamer. https://adealm01.de.festo.net/cb/tracker/1386233, 03 2019.

[5] Martin Hautzendorfer. Festo Automation Suite - Codebeamer. https://adealm01.de.festo.net/cb/tracker/598247, 02 2023.

[6] Robert Mueller. IT-Security Zitat. https://archives.fbi.gov/archives/news/speeches/combating-threats-in-the-cyber-world-outsmarting-terrorists-hackers-and-spies, 03 2012.

[7] Festo SEundCoKG. Servoantriebsregler CMMT-ST online kaufen | Festo DE. https://www.festo.com/de/de/p/servoantriebsregler-id_CMMT_ST/?q=CMMT-ST%7E%3Afesto-SortOrderScored, 2024.

Einsatz von Maschinellem Lernen zur Verbesserung der Klassifikation von Kundendokumenten

Tom Dinkelacker **Steffen Schober**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Postbeamtenkrankenkasse Körperschaft des öffentlichen Rechts, Stuttgart

Einleitung

Durch die ständigen wissenschaftlichen Fortschritte im Bereich der künstlichen Intelligenz (KI) bzw. des Machine Learning (ML) spielt die Integration dieser Technologien für immer mehr Unternehmen eine zunehmend größere Rolle. Insbesondere die Entwicklungen der letzten Jahre im Bereich des Natural Language Processing (NLP) bieten neue Möglichkeiten manuelle Prozesse umzudenken und repetitive Aufgaben mithilfe von Anwendungen zu erleichtern. NLP ist ein Bereich der künstlichen Intelligenz, welcher sich mit der Verarbeitung, dem Verstehen und der Generierung von natürlicher Sprache befasst. Durch NLP und zusammenhängende Technologien wie der Transformer-Architektur lassen sich komplexere Daten in natürlicher Sprache wie z.B. Texte nutzen, um innovative Lösungen zu erschaffen.

Zielsetzung

Diese Bachelorarbeit zielt darauf ab, ein konzeptionelles System zur Klassifikation von Kundendokumenten, insbesondere Freitextschreiben, zu entwickeln und eine fundierte Grundlage dessen zu schaffen. Dieses System soll mithilfe von Methoden des Machine Learning (ML) und Natural Language Processing (NLP) die Automatisierung der manuellen Klassifikation voranbringen und die gesamte Prozesslaufzeit verringern. In diesem Zusammenhang werden gängige Technologien betrachtet und miteinander verglichen, um ein bestmögliches Ergebnis zu erzielen. Insbesondere die Herausforderung der Datenaufbereitung sowie das Training, die Optimierung und der Vergleich verschiedener Transformer-Modelle werden ergründet. Dabei wird sowohl auf die theoretischen Grundlagen als auch die praktische Implementierung eingegangen. Der Prozess der Entwicklung eines produktionsfähigen Systems im Vergleich zu einem konzeptionellen System wird mit entsprechenden Herausforderungen bezüglich wichtiger Metriken wie Zuverlässigkeit und Geschwindigkeit dabei gesondert adressiert.

Transformer-Architektur

Attention is all you need. So lautet der Name des Papers von Vaswani et al. welches das Natural Language Processing (NLP) und viele andere Bereiche des Machine Learning seit dem Jahr 2017 grundlegend verändert hat. Erstmals wurde die Transfomer-Architektur und der „Self-Attention"-Mechanismus vorgestellt und angewandt, um die Verarbeitung von Sequenzdaten zu revolutionieren. Vor dem Nutzen von Transformern wurden Sequenzdaten in NLP-Problemen meist mit Recurrent Neural Networks (RNNs) verarbeitet. Diese hatten jedoch besonders Schwächen mit Geschwindigkeit und Verknüpfung von Informationen. [2]

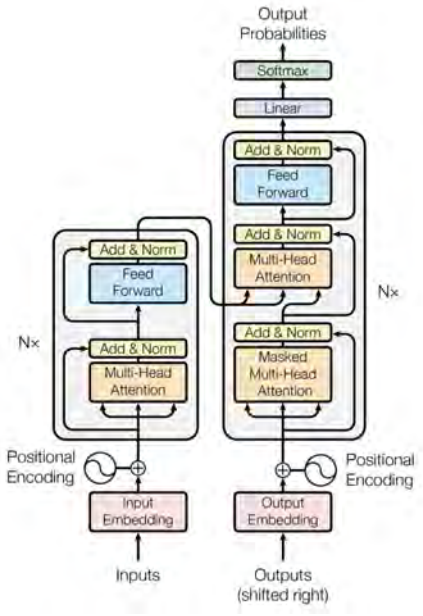


Abb. 1: Die Transformer Architektur [2]

Die Architektur eines Transformers besteht aus zwei Teilen: dem Encoder und dem Decoder (siehe Abbildung 1). Beide nutzen das Prinzip der Self-Attention, welche es dem Modell ermöglicht, Zusammenhänge zwischen den Input-Tokens (z.B. Worte) einer Sequenz (z.B. Text) zu verstehen. Dabei hilft Multi-Head-Attention mehrere solcher Zusammenhänge gleichzeitig zu betrachten, um das Verständnis des Kontexts zu erhöhen. Transformer verwenden darüber hinaus auch Position Encoding, um die Reihenfolge der Tokens in einer Sequenz zu berücksichtigen, und einen Softmax-Mechanismus, um die Vorhersage des nächsten Tokens in eine Wahrscheinlichkeit umzuwandeln. [2]
Durch diese Mechanismen hat die Transformer-Architektur maßgeblich dazu beigetragen, NLP-Anwendungen wie Textklassifikation, maschinelle Übersetzung und Sprachgenerierung zu verbessern.

Ausgangssituation

Für eine erste Umsetzung wird der Prozess der Nachklassifikation von Rückläufern genauer betrachtet. Bei Rückläufern handelt es sich um Vorgänge mit eingesendeten Freitextschreiben (z.B. E-Mails und Anschreiben) und angehängten Dokumenten (z.B. Rechnungen) von Kunden, welche in einem regelbasierten Klassifikationsverfahren falsch klassifiziert und zurückgewiesen wurden. Diese Vorgänge müssen daraufhin manuell bearbeitet und klassifiziert werden (siehe Prozess in Abbildung 2).

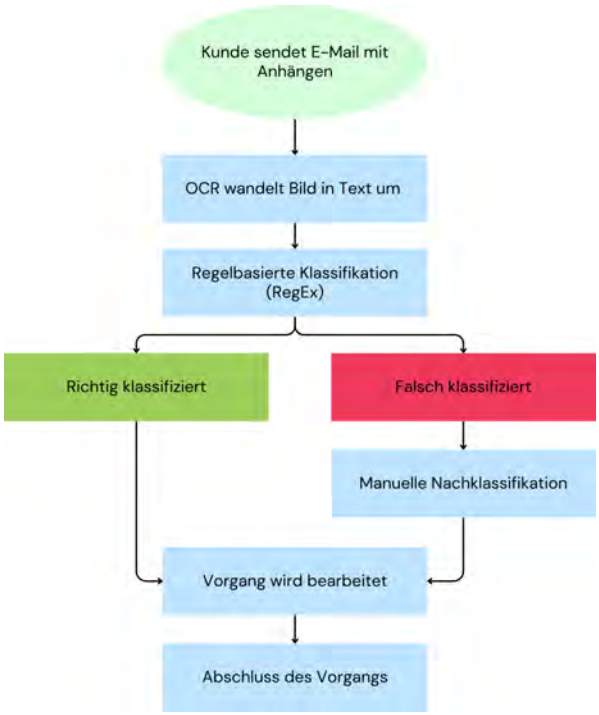


Abb. 2: Gesamtprozess - von der E-Mail zum abgeschlossenen Vorgang [1]

Die einzelnen Dokumente eines Vorgangs können sowohl wichtige Informationen wie z.B. Rechnungen oder Kündigungen als auch unwichtige Informationen enthalten. Besonders aus dem Anschreiben können aktuell nur wenige Informationen extrahiert werden. Aus dem Kontext aller Dokumente wird der Vorgang in der Regel einem von ca. 130 Postkörben zugewiesen und damit klassifiziert.

Umsetzung

Die Entwicklung stützt sich auf die Analyse bestehender digitaler Daten, darunter Kundendokumente als Bilddateien, Dokumententypen und manuelle Klassifikationsentscheidungen.
Die zentralen Schritte in der Umsetzung umfassen:

- Beschaffung, Extraktion und Aufbereitung relevanter Daten,
- Anwendung von NLP-Techniken zur Verarbeitung von Texten,
- Entwicklung und Training von einem oder mehreren ML-Modellen zur Dokumentklassifikation,
- Evaluierung der Ergebnisse hinsichtlich Genauigkeit, Effizienz und Einsetzbarkeit in einer Produktivumgebung.

Ein jedes Projekt im ML lebt von der Anzahl und Qualität der zur Verfügung stehenden Daten. Deshalb muss ein besonderes Augenmerk auf die Aufbereitung dieser gelegt werden. Insbesondere Datenschutzanforderungen und obsolete Daten wie E-Mail-Header, Anreden, Verabschiedungen und Namen müssen dabei entfernt werden. Ein trainiertes ML-Modell darf in personenbezogenen Daten keine Zusammenhänge sehen.
Mit den aufbereiteten Daten können so verschiedene NLP-Technologien erprobt und verschiedene Modelle trainiert und evaluiert werden. So kann z.B. neben einem Standard-Transformer wie BERT auch mit Document-Level-Attention (DLA) und einem Hierarchical Attention Network (HAN) gearbeitet werden. DLA erweitert den klassischen Attention-Mechanismus, indem es nicht nur lokale Zusammenhänge innerhalb eines Abschnitts, sondern auch globale Zusammenhänge über Abschnitte hinweg modelliert. Dies ermöglicht es, komplexe Informationen in umfangreichen Texten besser zu verstehen und in Kontext zu setzen. [3]

Cross-Document Attention

Neben Freitext-Anschreiben enthalten Einsendungen von Kunden oft auch weitere Dokumente. Diese Dokumente stehen in der Regel in einem direkten Kontext zueinander. Cross-Document Attention ist dabei ein spezieller Mechanismus, um Abhängigkeiten und

Beziehungen zwischen mehreren Dokumenten zu modellieren. Während klassische Attention-Mechanismen wie Self-Attention primär auf einzelne Dokumente angewendet werden, erweitert Cross-Document Attention den Kontext. Dabei werden einzelne Dokumente zunächst separat enkodiert und anschließend mithilfe von Cross-Attention-Mechanismen miteinander verglichen, um passende Verknüpfungen zu erzeugen. Dieser Ansatz ermöglicht es, Informationen aus verschiedenen Dokumenten zu kombinieren und Zusammenhänge zu verstehen. [4]
Im weiteren Verlauf könnte diese Technologie auch dazu beitragen, komplexe Kundendaten und ältere Dokumente miteinzubeziehen, um den Kontext und das Verständnis zu erhöhen.

Ausblick

Erste Ergebnisse zu Performance und Trefferquote der verschiedenen Technologien und Modelle bei der Klassifikation der Eingangsdokumente wirken vielversprechend und überliegen einer zufälligen Klassifikation. Jedoch gibt es noch einige Herausforderungen auf dem Weg bis zu einer produktionsreifen Verwendung jener Modelle. Während eine hohe Trefferquote bei der Klassifikation zwar wichtig ist, spielt auch die Treffersicherheit eine große Rolle. Um eine Teilautomatisierung anzustreben wäre z.B. denkbar nur jene klassifizierten Dokumente zu verwenden, in deren Richtigkeit sich ein oder mehrere Modelle sehr sicher sind. Auch bei falschen Klassifikationen gibt das Modell häufig eine hohe Treffersicherheit an, was die Zuverlässigkeit einschränkt.
Durch das weitere Training, aber vor allem die Verwendung von mehr und besser aufbereiteten Daten sowie der Verwendung von neuen Technologien, um vorhandene Daten besser zu nutzen, kann die Einsetzbarkeit des Systems gesteigert werden.
Die Ergebnisse sind ein wichtiger Schritt, die Ziele dieser Bachelorarbeit zu erreichen.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. https://arxiv.org/abs/1706.03762, 2017.

[3] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical Attention Networks for Document Classification. https://aclanthology.org/N16-1174, 2016.

[4] Xuhui Zhou, Nikolaos Pappas, and Noah A. Smith. Multilevel Text Alignment with Cross-Document Attention. https://doi.org/10.48550/arXiv.2010.01263, 2020.

Evaluierung einer Eventkamera zur Anwesenheitsdetektion von Objekten auf einer Mikrocontroller-Plattform

Thimo Dost **Clemens Klöck**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Leuze electronic GmbH + Co. KG, Owen

Eventkameras

Eventkameras gewinnen besonders in industriellen Bereichen immer mehr an Beliebtheit. Zu den Einsatzbereichen gehören Beispielsweise die Robotik, automatisiertes Fahren, die industriellen Automatisierung, die Pharmaindustrie und viele weitere. [3] Das liegt unter anderem daran, dass sie im Vergleich zu herkömmlichen Kameras eine höhere Dynamik bieten. Daten der Eventkamera werden asynchron generiert, das bedeutet, dass hier nicht zu fixen Zeitpunkten Events gesendet werden, sondern nur wenn eine Helligkeitsänderung im jeweiligen Pixel vorliegt. [2] Daraus ergeben sich folgende Vorteile:

- geringere Datenmengen
- ein geringerer Energieverbrauch
- eine bessere zeitliche Auflösung

Diese Eigenschaften machen es möglich, eine derartige Eventerkennung auf Mikrocontrollerbasis umzusetzen, auch wenn hier die verfügbaren Ressourcen geringer sind. In meiner Abschlussarbeit wird ein Eventsensor verwendet, dieser funktioniert nach dem selben Prinzip, wie eine Event Kamera.

Ziel der Abschlussarbeit

Das generelle Ziel der Abschlussarbeit ist es, die Performance eines Eventsensors auf einem Mikrocontroller im Hinblick auf Objekterkennung mit verschiedenen Algorithmen zu evaluieren. Mit jedem verwendeten Algorithmus finden mehrere Messungen statt. Hierfür werden unterschiedliche Akkumulationszeiten, also Zeitfenster indem Events des Sensors gesammelt und anschließend verarbeitet werden miteinander verglichen. Dadurch soll die Performance der Algorithmen analysiert werden. Es ist das Ziel, eine möglichst geringe Akkumulationszeit zu erreichen, mit der die Algorithmen gute Ergebnisse liefern. Die Ergebnisse werden Anhand der Ausführungszeit und der Detektionsrate betrachtet. Die Ausführungszeit zeigt, wie schnell der Algorithmus die Sensordaten verarbeiten kann. Die Detektionsrate hingegen veranschaulicht die Fähigkeit, relevante Merkmale der zu erkennende Objekte zu erfassen. Im weiteren Verlauf sollen mögliche Störeinflüsse untersucht werden und ob man diese durch das anpassen der Sensorparameter vermeiden kann.

Hardware

Die verwendeten Hardware Module wurden in folgendem Diagramm dargestellt.

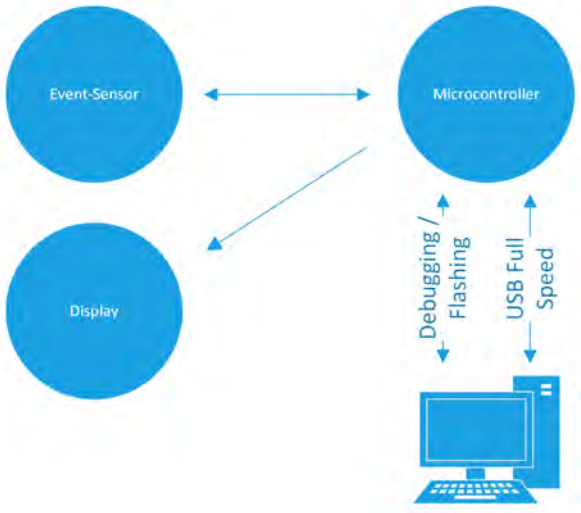


Abb. 1: Hardware Setup [1]

Der Eventsensor erfasst in Echtzeit die eingehenden Events und schreibt diese in den Speicher des Mikrocontrollers. Diese Daten werden anschließend in einem separaten Task ausgewertet und weiterverarbeitet. Für die Übertragung der Ergebnisse und Konfigurationsparameter ist eine separate USB-Schnittstelle implementiert. Das Display dient zur Darstellung der erfassten Events und eingestellten Parametern.

Vorgehen

Zu Beginn wurde ein neuer Task angelegt, indem die Verarbeitung der Eventdaten des Sensors stattfindet. Zur Visualisierung der eingehenden Events wurde eine Benutzeroberfläche auf dem Display angelegt. Diese wird zyklisch mit den Daten des Tasks aktualisiert. Für die Datenübertragung wurde eine geeignete Schnittstelle gesucht, die für die zuvor berechneten Datenmengen und Verarbeitungszeiten infrage kommt und zusätzlich in die Hardwarekonfiguration des Mikrocontrollers passt. Die Entscheidung viel hier auf eine Full-Speed USB Schnittstelle. Für die Finale Auswertung der Algorithmen, wurden folgende Modi implementiert:

- Aufnahme von Eventdaten zur Generierung eines wieder verwendbaren Datensatzes, der lokal auf dem Computer abgespeichert werden kann
- Übertragung des gespeicherten Datensatzes in den Mikrocontroller, sodass dieser mit verschiedenen Algorithmen ausgewertet werden kann
- Übertragung der Algorithmen Ergebnisse aus den erhaltenen Daten
- Verändern der Sensorparameter zur Laufzeit

Folgendes Diagramm stellt die verschiedenen Modi grafisch dar.

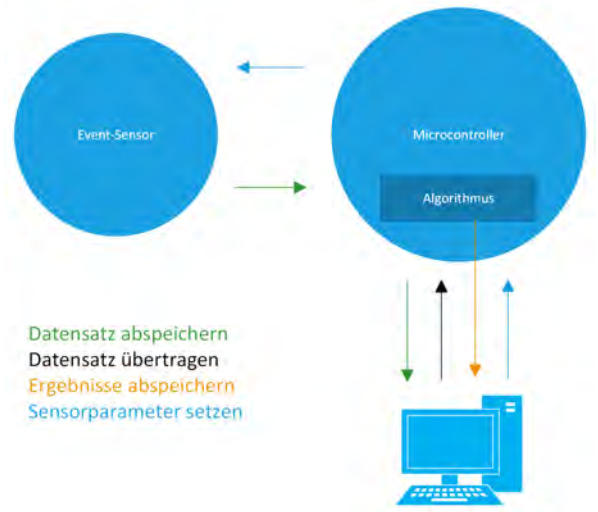


Abb. 2: verschiedene Modi [1]

Zur späteren Auswertung wurde mit Python ein Bedienoberfläche erstellt, mit der man Zugriff auf die Steuerung der verschiedenen Modi hat. Dies ermöglicht es, verschiedene Messungen mit unterschiedlichen Parametern durchzuführen, ohne die Applikation erneut auf den Mikrocontroller laden zu müssen. Die Auswertung der Algorithmen erfolgt ebenfalls mit Python.

Messungen/Messaufbau

Das Durchführen der Messungen findet auf einem, in 3 Achsen linear verfahrbaren Messtisch statt. Hierfür wird der Sensor, der Mikrocontroller und das Display in einer Halterung eingebaut und anschließend über dem Messtisch verfahren. Auf dem Tisch befinden sich mehrere Objekte, welche von dem Eventsensor detektiert werden sollen. Diese Messergebnisse werden ebenfalls am Computer ausgewertet.

Vorteile

Generell gibt es viele Möglichkeiten, eine derartige Objekterkennung umzusetzen. Ein großer Vorteil eines solchen Systems zeigt sich aber in der Erweiterbarkeit. Es kann mit geringem Aufwand an neue Szenarien angepasst werden und verspricht so eine lange Verwendbarkeit. Dies ermöglicht auch den Einsatz in verschiedenen Bereichen, was sich positiv auf anfallende Entwicklungsarbeiten und Kosten auswirkt.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jorg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based Vision: A Survey. https://arxiv.org/pdf/1904.08405, 2020.

[3] Maximilian Schenner. Eventkameras: Die Revolution der Fotografie? https://www.netzwoche.ch/news/2023-11-26/eventkameras-die-revolution-der-fotografie, 11 2023.

Design and Implementation of a Gateway for Controlling the State Machine of an Electric Drive via OPC UA FLC and Modbus TCP

Jason Patrick Duffy **Michael Scharf**

Department of Computer Science and Engineering, Esslingen University

Work carried out at Steinbeis Embedded Systems Technologies GmbH, Esslingen am Neckar

Introduction

The ongoing digitalization of industrial automation technology has significantly increased the need for communication between devices produced by disparate manufacturers. Historically, these manufacturers developed their own communication and information models. Consequently, the devices were only compatible with those from the same manufacturer. The design and implementation of interoperability between devices from different manufacturers was the responsibility of the application developer. This process was time-consuming and required in-depth knowledge of all the devices to be connected. OPC UA addresses this issue by offering universal information and communication models for these devices. The device manufacturer itself is responsible for implementing this. [6]
The OPC UA Field Level Communication (FLC) initiative seeks to extend the OPC UA standard to devices at the field level. This results in devices, spanning from basic sensors to sophisticated controllers, being able to communicate with one another even via cloud interfaces. [2] The resulting specifications are published under the designation OPC UA Field eXchange (UAFX). One of the components still in development is the UAFX Motion specification, which defines the standardized communication and information model for motion devices.
Within the specifications defined here, state machines play a pivotal role in the monitoring and controlling of electric drives. To test the feasibility of these newly defined state machines, a gateway shall be designed and implemented to control an electric drive using Modbus TCP and to connect the drive's own state machine with the UAFX Motion state machines. This gateway shall also be integrated into the UAFX Motion Prototyping Platform, which is an existing software platform used by members of the UAFX Motion working group to test their own implementations of the specification. The overall system structure is illustrated in Fig. 1.

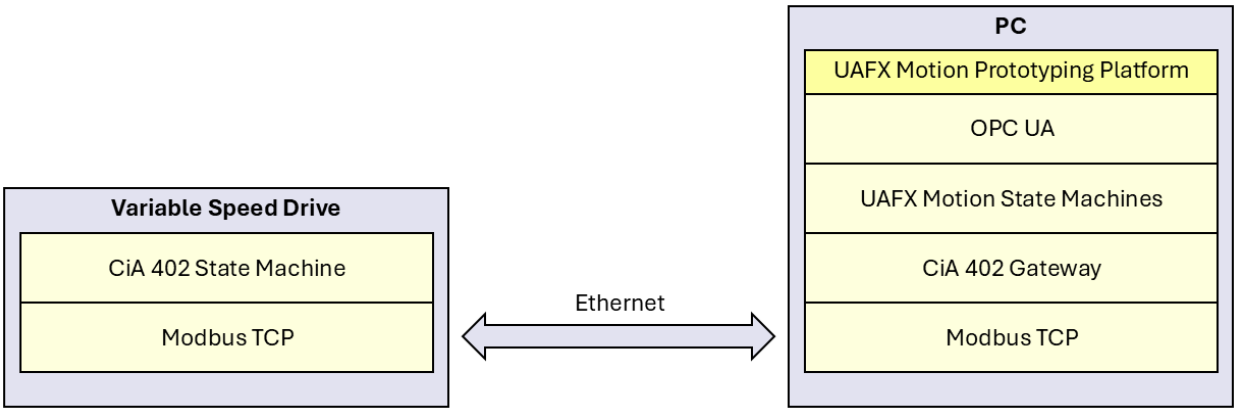


Fig. 1: System Structure Overview [7]

Modbus TCP

The Modbus protocol was developed in 1979 and facilitates communication between devices via a client-server architecture. Modbus is the most prevalent network protocol utilized in industrial production environments. [5]
Modbus is distinguished by a straightforward communication structure. A variety of function codes are defined that can be transmitted to a Modbus-compatible device to execute a standardized function. For example, function code '3' enables the sender to read the device's register contents. [4]
As Modbus is a protocol on the application level of the OSI reference model, communication is feasible via a multitude of underlying technologies. For this thesis, communication via TCP is of particular importance.

Electric Drive

In the context of this thesis, the electric drive is a Schneider Electric Altivar Process ATV630, which is a Variable Speed Drive (VSD) for synchronous and asynchronous motors. [9]
A VSD is a controller that regulates the speed of an electric motor based on automated inputs from an industrial process. By enabling precise adjustment of motor speed to meet process requirements, VSDs reduce energy consumption and ensure smoother operations. [10]
The ATV630 offers various communication and control options. Of particular significance for this thesis are its communication capabilities via Modbus TCP over Ethernet, enabling communication with a PC without the need for specialized equipment, and its support for controlling functionality through the standardized CANopen CiA 402 Power Drive System (PDS) state machine.
The ATV630 provides registers for reading and setting various variables, which can be accessed via Modbus TCP. This allows, for instance, for the current motor speed to be monitored. Additionally, these registers can be utilized to set the control word for the state machine and to read its status. [8]

State Machines

Two sides must be considered to control the drive: OPC UA and ATV630. The latter employs the CiA 402 state machine, which serves as the basis for the design. Both sides specify state machines explicitly for the control of electric drives.
The CiA 402 state machine is a highly flexible state machine that allows a great deal of freedom in the exact implementation on the manufacturer's side. Furthermore, the behavior of transitions between certain states can be defined using selectable presets, the implementation and existence of which depends on the specific implementation of the device manufacturer. This means their behavior may vary to some extent between devices developed by different manufacturers. This variation must be considered when designing the gateway. The CiA 402 state machine is controlled by a control word that encodes a variety of commands using a bit encoding system. [1]
The current concepts for the OPC UAFX Motion specification define a total of four state machines: Application, Fault, Input Converter and Axis. These state machines are interconnected to represent the functionality of a motion device. The Axis state machine is of particular significance in the context of drive control, as it precisely sets the behavior of a connected drive. In contrast to the CiA 402 state machine, the Axis state machine is not controlled by the setting of a single control word, but rather by modifying multiple variables within OPC UA. A control word regulates whether the drive should be in a running state or decelerating. [3]

Design of the Gateway

A link must be established between the Axis state machine and the CiA 402 state machine but a one-to-one link is not possible. For this to be feasible, the number of states would have to be identical, which is not the case. Furthermore, some actions that are resolved in the CiA 402 state machine through multiple transitions and states are executed in a single state and transition in the Axis state machine.
As a result, a more detailed mapping strategy was devised to link the two state machines. For this purpose, several application scenarios were defined, and sequence diagrams based on these scenarios were created to ascertain the data exchange and synchronization of the state machines. One such diagram, shown in Fig. 2, depicts a coast stop – a situation in which the drive comes to a standstill without active braking.
The gateway is designed to support drives compliant with the CiA 402 standard. To address variations in manufacturer's implementations, a generic interface is employed, which gets implemented on a per-device basis. Serving as a translation layer, the gateway bridges the Axis state machine and the CiA 402 state machine by evaluating and translating actions between the two. This facilitates basic velocity control of the drive via OPC UA.
Testing is conducted by simulating each defined application scenario to verify the completeness of the implementation.

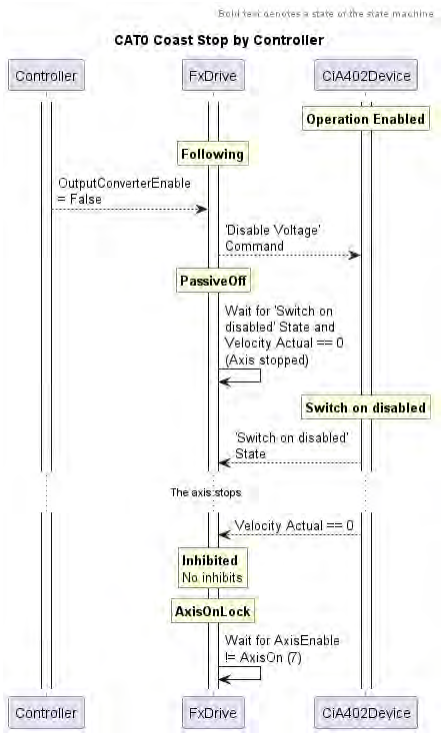


Fig. 2: Sequence Diagram for the Coast Stop Process [7]

Conclusion and Outlook

The development of this gateway represents a significant step in testing the UAFX Motion specification by integrating a real electric drive with the CiA 402 state machine. As the first implementation of its kind, it lays the foundation for evaluating the specification's completeness and feasibility in real-world scenarios. By enabling early detection of potential challenges, it helps reduce future time investment. Looking ahead, the gateway could be extended to support other CiA 402-compatible devices and accommodate additional state machines used in industrial automation.

References and figures

[1] CAN in Automation eV. CiA Final Work Draft 402 CANopen Drives and motion control device profile Part 2: Operation modes and application data Version 2.1.10. https://www.can-cia.org/can-knowledge/cia-402-series-canopen-device-profile-for-drives-and-motion-control/, 08 2006.

[2] OPC Foundation. Extending OPC UA to the field: OPC UA for Field eXchange (FX). https://opcfoundation.org/wp-content/uploads/2023/11/OPCF-FLC-Technical-Paper-C2C-EN.pdf, 11 2023.

[3] OPC Foundation. OPC Unified Architecture Field eXchange (UAFX) UAFX Motion - State Machines Whitepaper 2.0. https://reference.opcfoundation.org/, 05 2024.

[4] Modbus Organization Inc. Modbus Application Protocol Specification V1.1b3. https://modbus.org/docs/Modbus_Application_Protocol_V1_1b3.pdf, 04 2012.

[5] Modbus Organization Inc. Modbus FAQ: About The Modbus Organization. https://modbus.org/faq.php, 2024.

[6] Wolfgang Mahnke, Stefan-Helmut Leitner, and Matthias Damm. *OPC Unified Architecture*. Springer Berlin / Heidelberg, 2009.

[7] Own representation.

[8] Schneider Electric SE. Altivar Process ATV600 - Variable Speed Drives for Asynchronous and Synchronous Motors - EthernetIP Modbus TCP Manual: VW3A3720, VW3A3721. https://www.se.com/de/de/download/document/EAV64328/, 10 2017.

[9] Schneider Electric SE. Variable Speed Drive ATV630. https://www.se.com/uk/en/product/ATV630U22N4/variable-speed-drive-atv630-2-2kw-3hp-380-480v-ip21-ul-type-1/, 2024.

[10] David William Spitzer. *Variable Speed Drives: Principles and Applications for Energy Cost Savings*. Momentum Press, 4 edition, 2012.

Transformation von Security Operations Center durch Prozessoptimierung und Künstliche Intelligenz

Yasin Eraslan **Dirk Hesse**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma HBSN GmbH, Hornburg

Motivation und Problemstellung

Die zunehmende Menge und Komplexität von Sicherheitsdaten stellt Security Operations Center (SOCs) vor neue Herausforderungen. Die Bedrohungslandschaft ist stetig im Wandel und es besteht die Notwendigkeit, Bedrohungen schneller und präziser zu erkennen, zu analysieren und auf sie zu reagieren. Traditionelle SOC-Tools und -Technologien sind dazu nicht in der Lage, aufgrund hoher False-Positive Raten und mangelnder Skalierbarkeit. Durch den Einsatz von künstlicher Intelligenz (KI) durch Hacker werden Angriffe raffinierter und skalierbarer [7]. Der Lagebericht 2023 des Bundesamts für Sicherheit in der Informationstechnik (BSI) zeigt auf, dass täglich knapp 70 neue Schwachstellen in Softwareprodukten erfasst werden, mit steigender Tendenz. Dabei wird jede sechste Schwachstelle als kritisch eingestuft. Das BSI betont, dass KI neue Risiken schafft und Phishing-Mails schwerer zu identifizieren sind. Die Bedrohungen im Cyberraum werden vom BSI als so hoch wie nie zuvor eingeschätzt [2]. Die Einschätzung des BSI zur Bedrohungslage bestätigt die Ergebnisse der Bitkom-Studie „Wirtschaftsschutz 2024“. Durch Cyberangriffe fühlen sich zwei Drittel der Unternehmen in ihrer Existenz bedroht und acht von zehn Unternehmen geben an, von Datendiebstahl, Spionage oder Sabotage betroffen gewesen zu sein. Der Schaden für die deutsche Wirtschaft beläuft sich für das Jahr 2024 auf 267 Mrd. Euro [3]. KI bietet jedoch nicht nur Herausforderungen, sondern auch neue Chancen. Security-Anbieter versprechen, dass durch die Integration von KI in Security-Anwendungen Anomalien schneller erkannt werden, wodurch Maßnahmen zur Eindämmung schneller abgeleitet werden können, insbesondere im Vergleich zu traditionellen regelbasierten Systemen.

Ziele

In dieser Bachelorarbeit wird für die Serviceline SOC der Firma HBSN GmbH untersucht, ob der Einsatz von KI-Technologien in der bestehenden SIEM-Anwendung nicht nur möglich, sondern auch sinnvoll ist, um den steigenden Anforderungen gerecht zu werden. Dabei sollen die potenziellen Vorteile und Herausforderungen sowie Risiken analysiert und potenzielle Veränderungen in der Organisationsstruktur, den Rollen oder Verantwortlichkeiten berücksichtigt werden. Darüber hinaus werden Empfehlungen zu Prozessverbesserungen ausgesprochen, die durch die Analyse der gegenwärtigen Situation der Serviceline SOC resultieren. Im Rahmen dieser Bachelorarbeit wird eine qualitative Literaturanalyse durchgeführt. Experteninterviews werden zusätzlich zur Unterstützung der Forschungsmethode herangezogen.

IT-Sicherheit

Die IT-Sicherheit ist ein Teil der Informationssicherheit und umfasst ausschließlich Informationen, die mit Informationstechnologie verarbeitet werden. Die Informationssicherheit umfasst alle Arten von Informationen, einschließlich analoger Informationen. Ein weiterer Teilbereich der Informationssicherheit ist der Datenschutz, der ausschließlich den Schutz von personenbezogenen Daten umfasst. In den drei Bereichen Informationssicherheit, Datenschutz und IT-Sicherheit existieren zahlreiche Überschneidungen, überlappende Bedrohungsszenarien und Schutzmaßnahmen [5]. In Abbildung 1 sind die einzelnen Bereiche mit ihren Überschneidungen dargestellt.

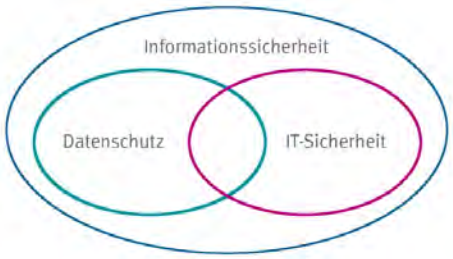


Abb. 1: Informationssicherheit [5]

Aufgaben und Ziele von SOCs

SOCs haben das Ziel, die Erkennung, Reaktion und Abwehr von Bedrohungen in Unternehmen zu verbessern. Bei einem SOC handelt es sich um ein internes oder externes Team von IT-Sicherheitsexperten. Ein SOC überwacht rund um die Uhr die gesamte IT-Infrastruktur. Kommt es zu einem sicherheitsrelevanten Ereignis, werden die IT-Sicherheitsexperten im SOC-Team benachrichtigt. Die Aufgabe des SOC-Teams besteht darin, Sicherheitsvorfälle zu erkennen, zu analysieren und darauf zu reagieren. SOCs wählen, betreiben und unterhalten IT-Sicherheits-Technologien und analysieren Bedrohungsdaten mit dem Ziel, die Sicherheitslage des Unternehmens zu verbessern. Durch den Einsatz eines SOC können Unternehmen Sicherheitssysteme vereinheitlichen und koordinieren, einschließlich Praktiken, Prozesse sowie die Reaktion auf Sicherheitsvorfälle. Dies resultiert typischerweise in verbesserten Präventionsmaßnahmen und Sicherheitsrichtlinien und ermöglicht eine schnellere Erkennung von Bedrohungen und eine effizientere Reaktion auf Sicherheitsbedrohungen. Durch ein SOC kann das Vertrauen der Kunden gestärkt werden und die Einhaltung von Datenschutzbestimmungen und Regularien für Unternehmen vereinfacht werden [1].

Abb. 2: SOC [4]

Technologien und Tools im SOC

In einem SOC kommen verschiedene Technologien und Tools zur Überwachung und zum Schutz der Unternehmensinfrastruktur vor Bedrohungen zum Einsatz. Eines der wichtigsten Tools, die ein SOC einsetzt, ist eine SIEM-Lösung, die Daten aus verschiedenen Protokollquellen aggregiert. Firewalls überwachen den Datenverkehr innerhalb der Unternehmensinfrastruktur sowie zu und aus dem Netzwerk. Der Datenverkehr kann auf Basis der vom SOC definierten Firewall-Regeln zugelassen oder blockiert werden. Durch den Einsatz von EDR-Technologien können Endpunkte geschützt und Bedrohungen abgewehrt werden [6].

Umsetzung

Im Rahmen der Umsetzung wurden die Funktionen und Produkte des aktuell eingesetzten SIEM-Tools analysiert, um potenzielle Möglichkeiten für die Integration von KI-Technologien zu identifizieren. Dabei wurde die geeignetste Option ausgewählt und einer detaillierten Analyse unterzogen. Auf Basis dieser Analyse wurden die potenziellen Vorteile und Nachteile der Integration abgeleitet. Anschließend wurde eine Roadmap zur schrittweisen Implementierung ausgearbeitet, die die erforderlichen Schritte, die benötigten Ressourcen und die zeitlichen Rahmenbedingungen beschreibt. Ein exemplarischer Use Case wurde durch ein Python-Programm teilautomatisiert, um den Nutzen einer Prozessoptimierung aufzuzeigen. Das Programm liest Protokolldateien aus einer Firewall ein und führt automatisierte Abfragen bei verschiedenen Threat-Intelligence-Tools durch. Dies ermöglicht eine effiziente Sammlung relevanter Informationen zu potenziellen Bedrohungen. Abschließend erstellt das Programm eine Beschreibung des Vorfalls sowie Handlungsempfehlungen für das SOC-Team. Damit wird die manuelle Arbeit reduziert und die Bearbeitungszeit für Sicherheitsvorfälle verkürzt.

Ausblick

Die Ergebnisse dieser Arbeit verdeutlichen, dass die automatisierte Analyse von Protokolldaten und die Integration von KI-Technologien in die bestehende SIEM-Anwendung der Serviceline SOC ein erhebliches Potenzial zur Optimierung bieten. Um die theoretischen Erkenntnisse praktisch zu validieren, ist die Durchführung eines Proof of Concept (PoC) erforderlich. Dieser PoC soll die technische Machbarkeit und den Mehrwert einer KI-Erweiterung im SIEM unter realen Bedingungen untersuchen.

Literatur und Abbildungen

[1] I BM. Was ist ein Security Operations Center (SOC)? https://www.ibm.com/de-de/topics/security-operations-center, 2024.

[2] Bundesamt für Sicherheit in der Informationstechnik. Die Lage der IT-Sicherheit in Deutschland. https://www.bsi.bund.de/DE/Service-Navi/Publikationen/Lagebericht/lagebericht_node.html, 01 2023.

[3] Bundesamt für Verfassungsschutz. Vorstellung der Bitkom-Studie „Wirtschaftsschutz 2024“. https://www.verfassungsschutz.de/SharedDocs/kurzmeldungen/DE/2024/2024-08-28-studie-bitkom.html, 08 2024.

[4] WBS IT-Service. Welche Vorteile bietet Ihnen ein Security Operations Center? https://www.wbs-it.de/services/soc, 01 2024.

[5] Universität Münster. INFORMATION SECURITY, IT-SECURITY AND DATA PROTECTION. https://www.unimuenster.de/Informationssicherheit/en/erkennen/Wert_von_Informationen_und_Schutzziele.html, 09 2023.

[6] Microsoft Security. Was ist ein Security Operations Center (SOC)? https://www.microsoft.com/de-ch/security/business/security-101/what-is-a-security-operations-center-soc#:~:text=SOC%2DTools%20und%20%2DTechnologien, 2024.

[7] Olivier Vareilhes. Datenzentrierter Ansatz zur Anomalieerkennung. https://www.security-insider.de/herausforderungen-loesungen-security-operations-center-a-65b04b165e94b26777f91db86ac073d4/, 07 2024.

Erstellung einer Google Cloud Platform Foundation für das Management und IT-Beratungsunternehmen MHP

Benjamin Erkel **Mirko Sonntag**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma MHP Management- und IT-Beratung GmbH, Ludwigsburg

Einleitung und Problemstellung

Das Aufkommen der Cloud-Plattformen stellt einen tiefgreifenden Paradigmenwechsel in der Gestaltung betrieblicher IT-Infrastruktur dar. Die weltweite Verbreitung und Adaption von Cloud-Plattformen sind in der Industrie zu einem essenziellen Bestandteil der IT-Infrastruktur geworden. Die Cloud ermöglicht Firmen im Vergleich zu bisherigen on-premise Infrastrukturen, ihre IT-Infrastrukturen kosteneffizienter, skalierbarer, flexibler und innovativer zu gestalten. Aufgrund der ständigen Verfügbarkeit einer Vielzahl von Services sind Organisationen in der Lage, komplette Systeme in kürzester Zeit zu entwickeln und auf den Markt zu bringen. Über die Cloud-Plattformen wird eine stetig steigende Anzahl an Services für dedizierte Kundenprobleme angeboten. Diese Entwicklung bietet Kunden das Potenzial, die benötigte Infrastruktur sehr gezielt aufzubauen und gleichzeitig Kosten zu reduzieren, indem nur Umfänge genutzt werden, die auch benötigt werden; das sogenannte „pay-as-you-go“. Firmen profitieren von der Vielzahl hoch entwickelten Sicherheitsservices, welche die Infrastruktur und die Daten schützen. In vielen technologischen Entwicklungen sind Cloud-Plattformen an der vordersten Front der Innovation. Aktuell zeigt sich dies in der Einführung neuer KI-Services. [7]

Einer der Top-3 führenden Cloud-Plattformen, gemessen an Marktanteil, ist die Google Cloud Platform, kurz GCP, wie es in der Abbildung 1 zusehen ist. Aufgrund der Historie von Google aus dem Suchmaschinen- und Werbegeschäft bietet die GCP eine robuste globale Netzwerkinfrastruktur. Ein besonderer Fokus liegt bei der GCP auf Angeboten aus den Bereichen Big Data und Datenanalyse sowie Künstliche Intelligenz und Machine Learning. [3]

Für eine große Organisation besteht die Herausforderung, Ordnung und Überblick auf der Plattform aufrechtzuerhalten. Um den Start auf der Cloud-Plattform für Entwicklerteams zu erleichtern und zu beschleunigen, ist es notwendig, Strukturen, Leitplanken und Richtlinien zu schaffen und zu überwachen. Ohne dies haben Entwicklerteams einen erschwerten Start.

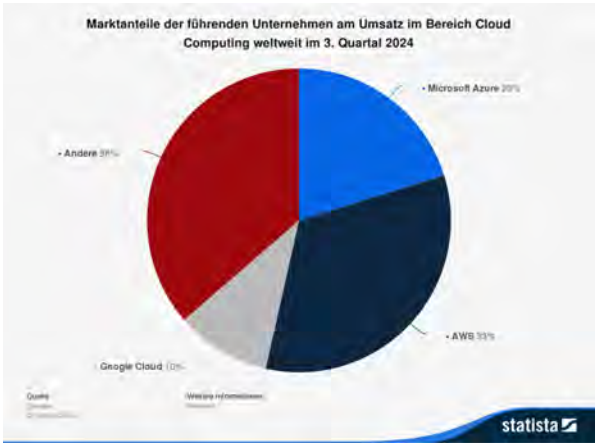


Abb. 1: Marktanteile der führenden Unternehmen am Umsatz im Bereich Cloud Computing weltweit im 3. Quartal 2024 [1]

Zielsetzung

Das Fundament für eine effiziente und nutzerfreundliche Cloud-Plattform innerhalb eines Unternehmens bildet eine Cloud-Foundation. Deshalb ist die Zielsetzung der Bachelorarbeit die Konzeptionierung und teilweise Erstellung einer Cloud Foundation auf der Google Cloud Platform, um die genannten Strukturen, Leitplanken und Richtlinien aufzubauen. Die Foundation soll eine Grundstruktur bieten, wodurch verschiedenste Projekte einer Organisation gemanagt und gefördert werden.

Damit eine Foundation regelmäßig, ohne tiefgreifenden manuellen Aufwand, an neu aufkommende Anforderungen angepasst werden kann, hat sich im Bereich der Cloud Foundations der Ansatz etabliert, diese mithilfe von Infrastructure-as-Code (IaC) aufzubauen. Die Struktur der Foundation wird mithilfe eines IaC

Tool definiert, welches es ermöglicht, per CI/CD-Pipeline Infrastruktur auf jeglicher Cloud-Plattform auszurollen (im Fall der Bachelorarbeit auf der GCP).

Cloud Foundation

Für die Migration auf die Cloud-Plattformen ist das bewährte Vorgehen in Unternehmen, eine Cloud Foundation zu erstellen. Sie stellt das Fundament innerhalb eines Unternehmens für die Arbeit in der Cloud dar und stellt der Organisation eine Grundstruktur bereit, um ihre Projekte auf die Cloud zu migrieren. Eine Cloud Foundation „umfasst die grundlegenden Komponenten, Ressourcen und Prozesse, die erforderlich sind, um eine sichere, skalierbare und effiziente Cloud-Umgebung aufzubauen." [5]
Um eine Foundation bauen zu können, muss identifiziert werden, was die grundlegenden Komponenten, Ressourcen und Prozesse einer Organisation sind. Diese grundlegenden Anforderungen können dann in Cloud-Governance-Richtlinien gefasst werden. Die Cloud-Governance-Richtlinien variieren stark zwischen verschiedenen Organisationen, da in jeder Organisation aufgrund unterschiedlicher Geschäftsmodelle und rechtlicher Bestimmungen andere Richtlinien und Strukturen existieren, aufgrund dieser Varianz ist eine eins-zu-eins-Kopie einer Cloud Foundation und der Cloud-Governance-Richtlinien in der Regel nicht möglich. Es ist jedoch möglich, Governance-Richtlinien aufgrund ihrer Funktion zu kategorisieren. Dies ist unabhängig von der Organisation möglich. Die Cloud Foundation Community hat diese in fünf funktionale Säulen eingeteilt. [4]

- Tenant Management: Entwicklung und Bereitstellung von Cloud-Umgebungen für Tenants, die Nutzer der Cloud-Plattform.
- Identity und Access Management (IAM): Managen von Identität und Zugriff auf die Cloud.
- Security & Compliance: Umsetzung und Regelung von Sicherheits- und Compliance-Richtlinien.
- Kostenmanagement: Kostenverwaltung und Kostennachverfolgung.
- Service-Ökosystem: Bereitstellung von verwalteten Diensten für die Entwicklerteams als Hilfestellung.

Beim Aufbau einer Foundation müssen nicht alle Säulen gleichzeitig erstellt werden. Manche Säulen der Cloud Foundation können zu einem späteren Zeitpunkt aufgebaut werden. Beispielsweise kann das Service-Ökosystem erst Blaupausen von anderen Projekten bereitstellen, nachdem die ersten Entwicklungsteams ihre Projekte gestartet haben.

Erstellung einer GCP Foundation mit Infrastructure-as-Code-Tool Terraform

Cloud-Provider bieten Organisationen, Werkzeuge und Ressourcen an, um erfolgreich auf ihrer Cloud durchzustarten Die GCP bietet eine Beispielfoundation an, welche in Terraform Code geschrieben wurde.
Terraform ist ein Infrastructure-as-Code-Tool, welches ermöglicht, per Code Ressourcen und Infrastrukturen in der Cloud aufzubauen. Terraform wurde von HashiCorp entwickelt. Damit ist es möglich, auf allen größeren Cloud-Plattformen Infrastrukturen zu erstellen. Es bietet mit der Terraform-Configuration-Language eine Konfigurationssprache, die Entwicklerinfrastruktur auf allen Cloud-Plattformen erstellen lässt. Dies hat den Vorteil, dass mit einem Werkzeug verschiedenste Infrastrukturen auf verschiedenen Plattformen erstellt werden können.
Ein weiterer Vorteil ist, dass der Terraform-Code in einem Git Repository gemanagten werden kann. Es ermöglicht eine übersichtliche Versionierung bei der Cloud Foundation-Entwicklung. Somit ist es möglich, viele Entwickler an der Foundation gleichzeitig arbeiten zu lassen. Veränderungen können manuell überprüft werden, bevor diese implementiert werden. Hierdurch wird die Codequalität und somit die Cloud-Architekturqualität abgesichert. Code-Änderungen und somit Anpassungen an der Cloudinfrastruktur werden dann automatisch durch eine Pipeline auf der GCP ausgerollt, wie in der Abbildung 2 zu sehen. Das Cloud-Foundation-Team kann so kontinuierlich an der Foundation entwickeln. Für den Bau der GCP Foundation werden mehrere Repositorys genutzt. Das entspricht dem Softwarearchitektur-Prinzip „Separation of Concern". Jedes Repository ist zuständig für Infrastrukturbereiche der GCP und orientiert sich grob an der erwähnten Säulenaufteilung. Die Struktur der Repositorys kann der Abbildung 3 entnommen werden. Momentan besteht die Architektur aus drei Repositorys:

1. bootsrap: Erstellt die CI/CD-Pipeline zwischen dem Git-Repository und der GCP. Die Pipeline wird dann immer ausgeführt, wenn Änderungen auf dieser oder auf anderen Pipelines entstehen.
2. org: Beinhaltet alle Organisation Infrastrukturen für geteilte Ressourcen, Netzwerke, Logging und IAM Policys.
3. environments: Erstellt drei Unterteilungen der Cloud-Umgebung (development, non-production, production)

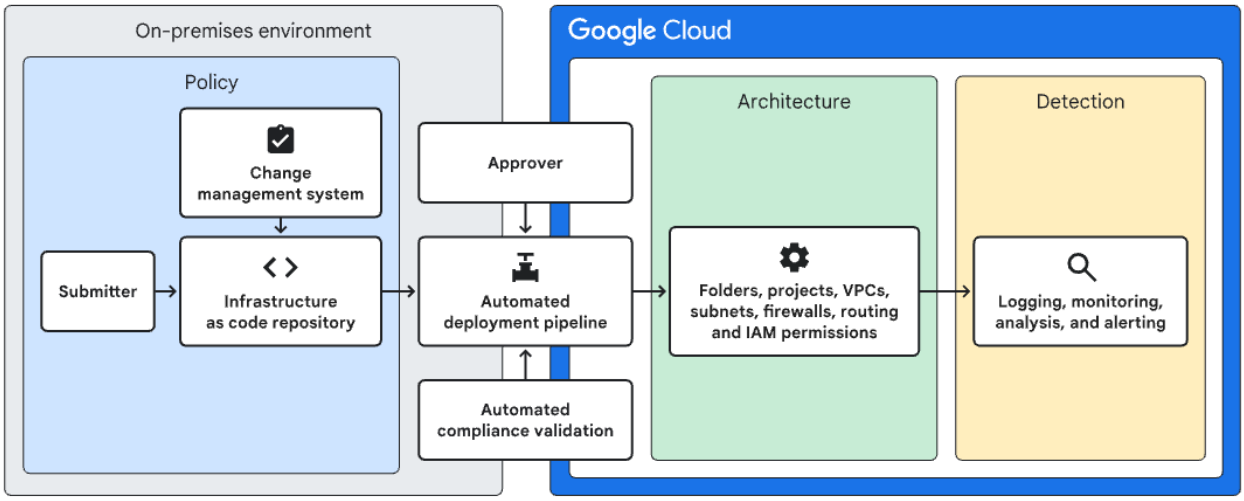


Abb. 2: End-to-End-Prozessarchitekture für Infrastrukturverwaltung in GCP [6]

Im Org Repository werden die ersten Säulen definiert. Mit den geteilten Ressourcen und den Netzwerken wird der Grundstein für das Service-Ökosystem gelegt. Die Netzwerke sind aber auch ein Bestandteil der Security und Compliance-Säule, da sie erste Netzwerk-Sicherheitsregeln beinhalten. Das Logging der Cloud ist ein Teil der Security und Compliance-Säule, da diese eine Übersicht der Cloud-Aktivitäten verschafft und somit unerwünschte Aktivitäten untersucht werden können. Dies ist auch Bestandteil der Kostenmanagement-Säule, da es ermöglicht, die Kosten den Verursachern zuzuordnen. Der Name der IAM Policys verrät bereits, dass sie zur IAM-Säule gehören. Das Environments Repository bildet mit seiner Orderstruktur eine Basis für die Entwicklerteams und ist somit ein Bestandteil der Tenant-Management-Säule, da sie eine Cloud-Umgebung für die Entwicklerteams bereitstellt.

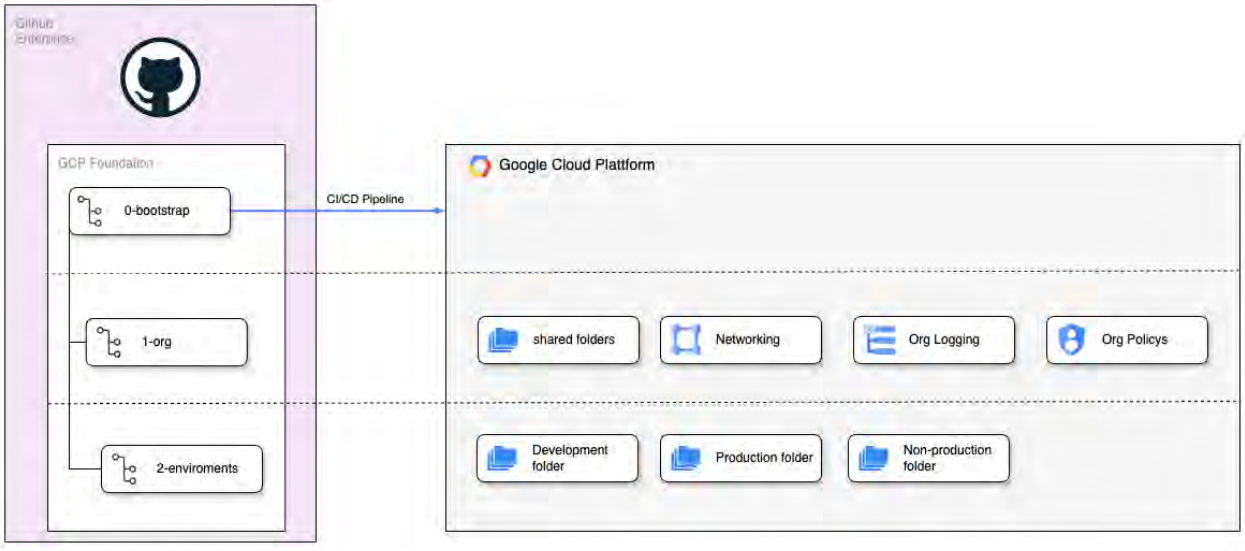


Abb. 3: Architekture Diagramm von der GCP Foundation [2]

Ausblick

Ein weiterer Teil der Bachelorarbeit wird die Erstellung einer Projekt-Factory sein. Die Projekt-Factory soll die Erstellung von Projekten auf der GCP vereinfachen und beschleunigen. Mithilfe einer CI/DC-Pipeline soll nach einem manuellen Approvel-Prozess auf der GCP ein neues Projekt mit vordefinierten Strukturen erstellt werden. Dieses Modul ist ein Bestandteil der Tenant-Management- und der Service-Ökosystems-Säule, indem es Projekte auf der Cloud managt und den Entwicklerteams eine Entwicklerumgebung zur Verfügung stellt.

Literatur und Abbildungen

[1] Canalys. Marktanteile der führenden Unternehmen am Umsatz im Bereich Cloud Computing weltweit im 3. Quartal 2024 [Graph]. https://de.statista.com/statistik/daten/studie/150979/umfrage/marktanteile-der-fuehrenden-unternehmen-im-bereich-cloud-computing/, 2024.

[2] Eigene Darstellung.

[3] Tobias Regenfuß and Timo Nink. Was kann die Google Cloud? https://www.cio.de/a/was-kann-die-google-cloud,3667814, 2024.

[4] Johannes Rudolph, Felix Zieger, et al. What is a Cloud Foundation. https://cloudfoundation.org/understanding-cloud-foundation/#why-build-a-cloud-foundation, 2023.

[5] Christian Scholz. Cloud Foundation. https://arc.net/l/quote/wlhscjyz, 2024.

[6] Ohne Verfasser. Blueprints für Unternehmensgrundlagen. https://cloud.google.com/architecture/security-foundations?hl=de, 2023.

[7] Ohne Verfasser. Vor- und Nachteile von Cloud-Computing. https://cloud.google.com/learn/advantages-of-cloud-computing?hl=de, 2024.

Präzise Linieninformationen durch Fusion von Kamera- und HD-Kartendaten zur Validierung von Fahrerassistenzsystemen

Karol Fedurko **Markus Enzweiler**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Robert Bosch GmbH, Leonberg

Einleitung

Zum Schutz von Menschenleben sind moderne Fahrzeuge im Straßenverkehr mit Fahrerassistenzsystemen ausgestattet. Insbesondere aktive Spurhalteassistenten und damit verbundene Sicherheitsfunktionen zielen darauf ab, sowohl den Fahrkomfort, als auch die Sicherheit aller Verkehrsteilnehmer zu erhöhen. Die Evaluierung der Funktionalität und des Verhaltens auf Verkehrssituationen dieser Systeme erfordert die Durchführung von Validierungsstrategien. Diese umfassen unter anderem reale Testfahrten. Dabei werden auf vordefinierten Referenzstrecken sowohl subjektive Eindrücke als auch objektive Daten gesammelt. Die Daten ermöglichen die Bestimmung von Schlüsselkennzahlen, welche wiederum Rückschlüsse auf die Leistungsfähigkeit der Systeme zulassen. Die Ausstattung eines Testfahrzeuges umfasst eine Videokamera, welche in periodischen Abständen Linieninformationen aus Sicht des Fahrzeuges aufzeichnet. Allerdings variiert die Qualität und Quantität der Linien in Abhängigkeit von extrinsischen sowie intrinsischen Einflüssen, weshalb die Ermittlung von verlässlichen Schlüsselkennzahlen zur Herausforderung wird. Deswegen befasst sich die Arbeit mit der Fusion von Linieninformationen auf Basis einer Kamera sowie mit HD-Kartendaten. Ziel ist es, präzise Linien zu bestimmen, um die Zuverlässigkeit der Schlüsselkennzahlen zu stärken.

Geodatenbank

Zunächst ist ein zentrales System zur Verwaltung der Linieninformationen notwendig. Insbesondere bedarf die Abfrage relevanter Daten eine effiziente Struktur, da bereits über zwei Millionen Punkte, zur Abbildung einer Karte mit lediglich 60 Kilometer Reichweite, erforderlich sind. Nach einer Evaluierung diverser Datenbankmanagementsysteme (DBMS) wurde die Entscheidung für eine PostGIS-Datenbank getroffen, welche als Erweiterung des relationalen DBMS PostgreSQL zur Verfügung steht. Die Erweiterung ermöglicht es, Daten von beliebigen Koordinatenreferenzsystemen (CRS) abzuspeichern. Darüber hinaus implementiert das DBMS sowohl Datentypen als auch Funktionen für geografische Daten, die dem ISO 19125-Standard entsprechen [4]. Um die für eine Testfahrt relevanten Punkte aus der Datenbank zu extrahieren, wird die Fahrzeugtrajektorie als offener Polygonzug in die Abfrage eingegeben. Die Datenbank kann nun mittels einer räumlichen Indexstruktur eine effiziente Filterung nach relevanten Punkten durchführen und die entsprechenden Ergebnisse wiedergeben. Die Aktualität der globalen Punkte beschreibt eine maßgebliche Fehlerquelle, insbesondere in Bezug auf die Plattentektonik. Die eurasische Platte bewegt sich jährlich etwa 2,5 Zentimeter von der nordamerikanischen weg [6]. Aus diesem Grund werden für CRS in regelmäßigen Abständen Korrekturen in Form von Epochen durchgeführt [10]. Die Datenbank verfügt über Methoden zur Transformation zwischen Koordinatenreferenzsystemen und Epochen, wodurch die Aktualität der Daten gewährleistet wird [4]. Innerhalb der Datenbank werden alle Punkte in ein einheitliches ellipsoidisches CRS, mit Polarkoordinaten, überführt. Die Abb. 1 zeigt einen Ausschnitt einer Karte (Linien in hellblau).

Abb. 1: Beispiel einer Linienkarte [3]

Clustering

Im Anschluss an die Extrahierung der relevanten Punkte erfolgt ein Clustering um zu bestimmen, welche Punkte fusioniert werden können. Um ein Clustering zu ermöglichen, ist zunächst eine Transformation der Polarkoordinaten in ein kartesisches Koordinatensystem erforderlich, wobei hier die Snyder-Methode (1987) zum Einsatz kommt. In der Literatur wurden zahlreiche Methoden für das Clustering von Linieninformationen vorgeschlagen. Eine Variante ist das Partitionsverfahren namens K-Means, bei dem k-Schwerpunkte initialisiert werden. Anschließend erfolgt eine iterative Zuordnung der Punkte mit Hilfe der Distanz zu einem Schwerpunkt. Der Schwerpunkt wird für jeden Schritt durch den Mittelwert aller Punkte im Cluster aktualisiert. Die Cluster können nicht überlappen und der Algorithmus funktioniert optimal für sphärisch angeordnete Daten, ist jedoch sehr anfällig für Verzerrung, aufgrund von Ausreißern [9]. Die Bestimmung des Parameters k stellt im Falle der Linieninformationen eine nicht triviale Herausforderung dar. Im Gegensatz zu Partitionsverfahren sind dichtebasierten Verfahren zu nennen, wie der bekannte Vertreter namens DBSCAN. Diese Methode bestimmt die Cluster, basierend auf der Masse der Punkte in Regionen. Insbesondere können die Cluster von beliebiger Form sein und Ausreißer identifiziert werden [5]. Dieser Ansatz erweist sich für das Clustering von Linieninformationen als robuster, allerdings weist er einen gravierenden Nachteil auf. Es ist nicht möglich, zwischen unterschiedlichen Datensätzen zu unterscheiden, was bei einer hohen Abtastrate der Datenbankpunkte zu unerwünschten Nebeneffekten führen kann. Aus diesem Grund wird ein Algorithmus implementiert, welcher für jeden relevanten Datenbankpunkt basierend auf der euklidischen Distanz neue Beobachtungspunkte zuordnet (Siehe Abb. 2).

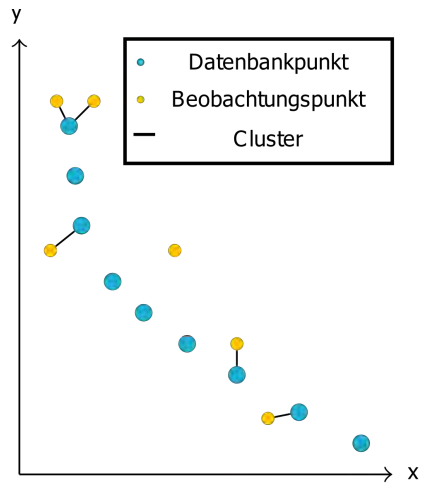


Abb. 2: Punktclustering [3]

Sensordatenfusion

Die Sensordatenfusion beschreibt die Verknüpfung von Daten mehrerer Quellen, mit dem Ziel, die Qualität der verfügbaren Daten zu maximieren und Fehler zu minimieren [7]. Im Rahmen der Anwendung auf die Linieninformationen ist eine Fusion von bereits vorliegender HD-Karten- und Videodaten sowie aktueller Testfahrtdaten, erforderlich. Daher vergleicht die Thesis verschiedene Fusionsmethoden, darunter:

1. Regularisierter Least-Square Polyfit

Die fundamentalste Fusionsmethode für die Problemstellung basiert auf einer Polynomregression. Hierbei wird ein Polynom variabler Ordnung durch die Punkte anhand eines Minimierungsproblems, hier „Least Squares“, approximiert. Eine Tikhonov-Regularisierung reduziert das Risiko einer Überanpassung, indem die Polynomkoeffizienten eingeschränkt werden [2]. Da Linienmarkierungen und Fahrbahnbeschränkungen in verschiedensten Formen vorkommen, eignen sich reguläre Polynome jedoch nicht.

2. B-Splines

Obwohl im Straßenbau, vor allem auf Autobahnen, die Modellierung mittels Klothoiden üblich ist, erweisen diese sich als suboptimal, um sämtliche Strukturen, insbesondere in ländlichen Gebieten, abzudecken. Eine Alternative, um komplexe Formen und Strukturen abzubilden, stellen B-Splines dar [1]. Hierbei handelt es sich um abschnittsweise definierte Polynome, die durch Knotenpunkte unter Berücksichtigung einer Stetigkeitsbedingung miteinander verbunden sind [2].

3. Maximum A Posteriori

Maximum A Posteriori (MAP) beschreibt eine Methode der Bayesschen Statistik. MAP erlaubt die Schätzung des maximalen Werts einer unbekannten Variable unter Berücksichtigung von a priori-Wissen und beobachteten Daten. Durch die Fusion der Daten erhält man a posteriori-Wissen [8].

Ergebnisse

Um die Fusionsmethoden effektiv vergleichen zu können wird eine vordefinierte Karte als Ground Truth (GT) definiert, welche künstlich mit normalverteiltem Rauschen sowohl in x als auch in y- Richtung zweifach modifiziert wird. Die Datenbankkarte weist ein geringeres Rauschen auf und simuliert eine höhere Sensorqualität. Die „beobachtete“ Karte hingegen zeigt einige Ausreißer. Die Fusion der zwei Karten zielt darauf ab, eine Karte zu generieren, die sich der Ground Truth annähert. Zur objektiven Auswertung werden einige Metriken herangezogen, unter anderem die max. Abweichung und die Standardabweichung. Der longitudinale Fehler ist durch die Abhängigkeit von der Fahrtrichtung irrelevant, weshalb zur Auswertung ein bahnfestes Koordinatensystem unabdingbar ist [11].

Die Fahrzeugtrajektorie $S = \{(x_i, y_i)|i = 0, ..., n-1\}$ dient als Referenzkurve. Jeder Punkt aus $\mathbf{p}_{GT}$ und $\mathbf{p}_{Vgl}$ wird auf die Referenzkurve näherungsweise orthogonal projiziert, um jeweils die laterale Abweichung $d_{GT,i}$ bzw. $d_{Vgl,i}$ mit $d_i = \min ||\mathbf{s}_i \cdot \mathbf{p}_i||$ zu bestimmen. Anschließend wird jedem GT-Punkt ein Vergleichspunkt im bahnfesten Koordinatensystem mit der minimalen euklidischen Norm, unter Berücksichtigung der Vorzeichen, zugeordnet. Der laterale Fehler e ergibt sich dann anhand Subtraktion der lateralen Abweichungen und dient als Grundlage für die Kalkulation der Metriken. Die Vorgehensweise wird in Abb. 3 dargestellt.

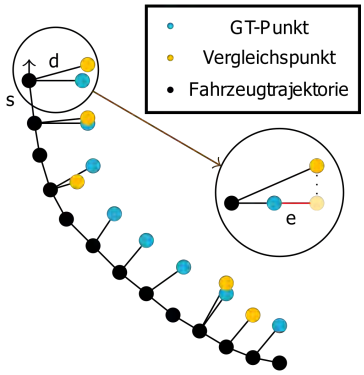


Abb. 3: Bahnfestes Koordinatensystem [3]

Ein Auszug der Resultate lässt sich wie folgt zusammenfassen: Während Prä-Fusion die maximale Abweichung 1,7 Meter und die Standardabweichung 0,08 Meter betrug, verbessert sich die maximale Abweichung auf 0,3 Meter, während die Standardabweichung auf 0,03 Meter sinkt. Diese Genauigkeit wird durch eine hybride Fusion von B-Splines und MAP erreicht. Im finalen Schritt erfolgt eine Aktualisierung der Linieninformationen innerhalb der Datenbank.

Schlussfolgerung

Die Fusion von bestehenden Kartendaten mit neuen Beobachtungsdaten, auch wenn diese von geringerer Qualität sind, führt zu einer Verbesserung der Genauigkeit der Linieninformationen. Insbesondere können lückenhafte Beobachtungsdaten durch einen zweiten Datensatz sinnvoll aufgefüllt werden. Dies führt zu einer zuverlässigeren Aussagekraft der Schlüsselkennzahlen bei der Bewertung von Fahrerassistenzsystemen und trägt zu einer höheren Sicherheit bei der Weiter- und Neuentwicklung dieser Systeme bei.

Literatur und Abbildungen

[1] A. Abramov, C. Bayer, C. Heller, and C. Loy. A flexible modeling approach for robust multi-lane road estimation. In *Intelligent Vehicles Symposium*, pages 1386–1392. IEEE, 2017.

[2] S. Boyd and L. Vandenberghe. *Introduction to Applied Linear Algebra – Vectors, Matrices and Least Squares*. Cambridge: Cambridge University Press, 2018.

[3] Eigene Darstellung.

[4] PostGIS Development Group. PostGIS 3.5.1 Manual. https://postgis.net/docs/manual-3.5/, 2024.

[5] M. Ester, H. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd, volume 96*, pages 226–231, 1996.

[6] National Geographic. Continental Drift. https://education.nationalgeographic.org/resource/continental-drift/, 2024.

[7] J. Kiebler. *Lokalisierung Und Fahrzustandsschätzung Für Eine Vollautomatisierte Elektrische Fahrzeugplattform*. Wiesbaden: Springer Fachmedien, 1 edition, 2024.

[8] M. Kumar, D. P. Garg, and R. A. Zachary. A generalized approach for inconsistency detection in data fusion from multiple sensors. In *2006 American Control Conference*, pages 2078–2083. IEEE, 2006.

[9] S. Lloyd. Least squares quantization in PCM. In *IEEE Transactions on Information Theory*, volume 28, pages 129–137. IEEE, 1982.

[10] Deutsches Institut für Normung. *Geoinformationen – Koordinatenreferenzsysteme (ISO 19111:2019)*. Berlin: Beuth Verlag GmbH, 2020.

[11] M. Werling. *Ein neues Konzept für die Trajektoriengenerierung und -stabilisierung in zeitkritischen Verkehrsszenarien*. Karlsruhe: KIT Scientific Publishing, 2011.

Analyse und Vergleich prototypischer HTMX- und Next.js-Anwendungen in der Kommunikation mit dem Payload CMS Framework

Robert-Bogdan Fesko **Harald Melcher**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma pep.digital GmbH, Esslingen am Neckar

Einleitung

In den letzten Jahre hat sich das Webumfeld signifikant verändert und erweitert. Modernen Anwendungen verlangen heutzutage nahtlose Performance und hohe Interaktivität, während es am Markt massenhaft verschiedene Technologien für unterschiedliche Zwecke gibt. Diese finden wiederum ihren Weg in die optimale Lösung, die dem Projekt des Stakeholders am besten gerecht wird. In dieser Hinsicht spielen Content-Management-Systeme (CMS) die entscheidende Rolle bei der Verwaltung von Inhalten für eine Website. Traditionelle CMS sind zwar bewährte Lösungen, aber für einen neuen Trend, das Headless CMS, findet jetzt großen Anklang unter den Entwicklern und Unternehmen. Zudem suchen ständig neue Konzerne nach innovativen Alternativen, die sie im digitalen Raum für den Ausbau ihrer Online-Präsenz nutzen können.

Zielsetzung

Diese Arbeit untersucht das Zusammenspiel von HTMX und dem Payload CMS im Vergleich zur etablierten Lösung mit Next.js. Durch die Entwicklung zweier prototypischer Blog-Anwendungen werden Backend-Integrationsmöglichkeiten, Dokumentation sowie Popularität analysiert. Um Objektivität zu gewährleisten, liegt der Fokus auf der Kernfunktionalität der Anwendungen. Die Ergebnisse sollen klare Entscheidungsgrundlagen für Entwickler sowie folgende Projekte bei der Auswahl passender Webtechnologien liefern.

Headless CMS

Bei den sogenannten "Coupled CMS"werden Frontend und Backend eng miteinander verknüpft. Die serverseitige Generierung von HTML-Inhalten erschwert die Nutzung für verschiedene Kanäle, beispielsweise Apps, Sprachassistenten oder Social Media, da die Inhalte nur mit großem Aufwand an die jeweiligen Anforderungen angepasst werden können. Im Gegensatz dazu erfolgt bei einem Headless CMS eine Trennung der Inhaltspflege von der Darstellung. Die Erstellung und Pflege der Inhalte erfolgt weiterhin in einer für die Nutzer leicht verständlichen Administrationsumgebung, jedoch nicht mehr fest an ein bestimmtes Frontend gebunden. Anstelle einer serverseitigen Generierung von HTML werden die Inhalte über flexible Application Programming Interfaces (APIs), typischerweise Representational State Transfer (REST) oder GraphQL, bereitgestellt. Die Entkopplung der Architektur erlaubt die plattformunabhängige Nutzung der Inhalte sowie eine vereinfachte Anpassung und Integration. Des Weiteren fördert ein Headless CMS die Nachhaltigkeit und Wertschöpfung des erstellten Inhalts, da eine problemlose Integration in neue oder bestehende Kanäle möglich ist. Die Vielzahl verfügbarer Headless-Lösungen, sowohl Cloud- als auch selbst gehostete Varianten, verdeutlicht das wachsende Interesse an dieser flexiblen, zukunftsorientierten Herangehensweise an das Content-Management. [5]

Abb. 1: Visualisierung eines Headless CMS [2]

Payload CMS mit Next.js

Das Content-Management-System (CMS) ist notabene die Notwendigkeit und Bedarfsfrage für den Ersteller von Inhalten, der eine einfache Handhabung der Daten erwartet. Mit einem solchen System können Benutzer ohne viele technische Kenntnisse Daten erstellen, ver-

walten und bearbeiten. Das Payload CMS Framework ist eine Open-Source CMS-Lösung, mit dem die Möglichkeit besteht, Daten in einer Datenbank zu speichern und sie über GraphQL- und REST-APIs zu manipulieren. Zudem können Daten über eine React-basierte Benutzeroberfläche einfach manipuliert werden. Im Gegensatz zu herkömmlichen monolithischen CMS-Lösungen trennt Payload das Content-Management absichtlich von der Präsentationsebene. Die Inhalte werden zentral über Schnittstellen bereitgestellt und können nahtlos in eine Vielzahl von Frontends und Technologien eingebunden werden. [4]

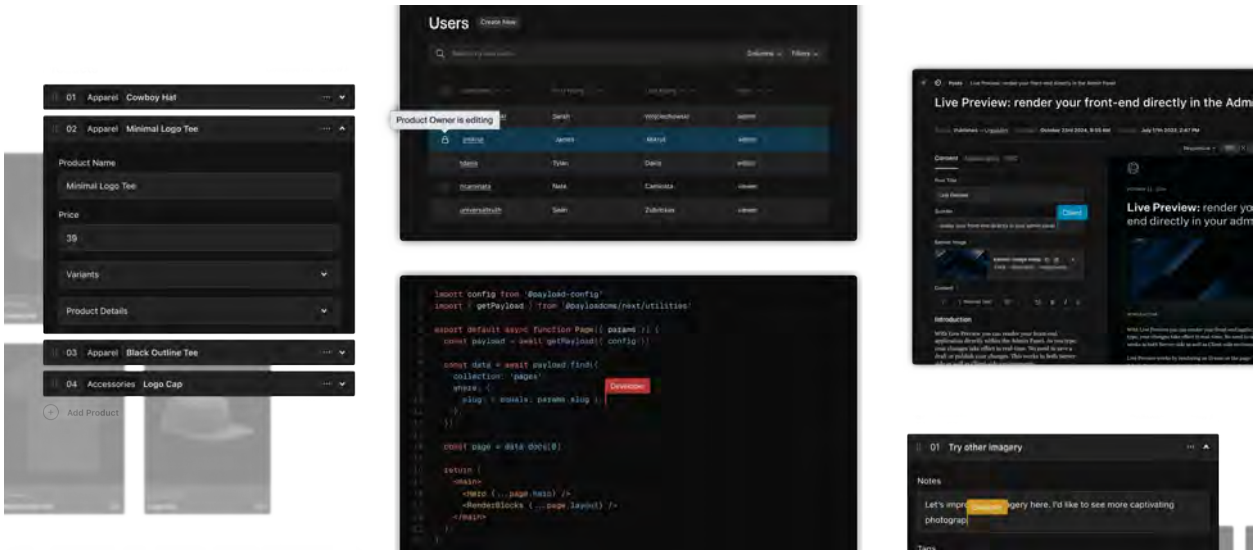


Abb. 2: Übersicht der Funktionen von Payload CMS [1]

Mit dem Übergang von der Beta- zur Release-Phase der dritten Version des CMS Payload ist es möglich, das System vollständig in Next.js zu integrieren, das zur Lieblingstechnologie der Entwickler dieses CMS-Systems geworden ist. [7] Dazu zählen unter anderem die Möglichkeit, direkt hinzugefügte Komponenten in die Oberfläche der Inhaltsverwaltung einzubauen, sowie die damit einhergehende Entlastung des Rendering-Prozesses auf der Client-Seite. [3]

HTMX Bibliothek

In den früheren Tagen der Webentwicklung gab es serverseitige Modelle von Skriptsprachen wie PHP, Ruby on Rails, ASP.NET usw., um den Prozess der Erstellung dynamischer Webseiten zu erleichtern. Mit dem Aufkommen von Frameworks, die das Rendering auf die Benutzer-Clients verlagern, wie React, Angular und Vue, ist eine neue Ära angebrochen, eine Ära mit höherer Interaktivität, aber auch mehr Komplexität. HTMX stellt eine zeitgemäße Lösung dar, die diese Kluft überbrückt. Sie ermöglicht es Entwicklern, interaktive Webschnittstellen in HTML mit einem Minimum an JavaScript zu erstellen. Der große Vorteil dabei ist die dynamische Aktualisierung von HTML-Teilen als Folge von Anfragen an den Server. HTMX hat den Aufwand für die Pflege und Entwicklung von Webanwendungen drastisch reduziert. Diese Methode ist eine vielversprechende Alternative zu rein clientseitigen Frameworks und stößt auf die Sympathie vieler Entwickler, die sich einfachere, aber dennoch leistungsfähige Ansätze für die moderne Webentwicklung wünschen. [6]

Vorgehensweise

Vor der eigentlichen Implementierung entstand zunächst eine grobe Gesamtstruktur des Systems. In diesem Zusammenhang wurden sowohl funktionale als auch nicht-funktionale Anforderungen definiert. Ein Systemdiagramm veranschaulichte anschließend die Beziehungen zwischen den verschiedenen Komponenten. Zusätzlich dienten Wireframes dazu, die grundlegende Gestaltung der Seiten beider Webanwendungen übersichtlich darzustellen.
Im Anschluss begann die konkrete Umsetzung der Anwendungen. Im Backend wurden mehrere Collections angelegt, die festlegen, welche Inhalte das Payload CMS automatisch über den Datenbankadapter in die PostgreSQL-Datenbank speichert. Der Adapter übersetzt interne Datenstrukturen nahtlos in native PostgreSQL-Formate. Darüber hinaus integrierte das Projekt eine Funktion, mit der definierte Komponenten (sogenannte Blöcke) in einem Rich-Text-Editor konfiguriert werden können. Beide Frontend-Anwendungen greifen über eine bereitgestellte API auf diese Collections zu.

Für die Anwendung, die auf die HTMX-Bibliothek setzt, wurde einen Rich-Text-zu-HTML-Konverter entwickelt. Damit lässt sich der HTML-Inhalt direkt per API-Aufruf beziehen und im Frontend darstellen. Die native Integration von Payload mit Next.js führte zu einer noch strukturierteren Inhaltspräsentation: Eine speziell definierte React-Komponente serialisiert die "Blöckeïn React-JSX-Elemente.

Ausblick

Die Umsetzung aller Anforderungen sowie die während des Implementierungsprozesses erworbenen Kenntnisse eröffnen die Möglichkeit, die Effizienz der Verwendung der HTMX-Bibliothek in der Kommunikation mit Payload CMS zu analysieren. Dabei werden Faktoren wie die Implementierung zusätzlicher Funktionalitäten im Vergleich zur nativen Integration mit Next.js sowie die Antworten auf die zu Beginn der Arbeit gestellten Forschungsfragen berücksichtigt.

Literatur und Abbildungen

[1] Payload CMS. Payload: The Next.js Headless CMS and App Framework. https://payloadcms.com/, 2024.

[2] Eigene Darstellung.

[3] Payload Dev. The best way to build a modern backend. https://payloadcms.com/developers, 2024.

[4] Payload Docs. What is Payload? https://payloadcms.com/docs/getting-started/what-is-payload, 2024.

[5] Kevin Erath. Was ist ein Headless CMS? https://pep-digital.de/blog/was-ist-ein-headless-cms, 01 2021.

[6] Alex Merced. What is HTMX? Why it Matters? and How to use it. https://dev.to/alexmercedcoder/what-is-htmx-why-it-matters-and-how-to-use-it-10h3, 12 2023.

[7] James Mikrut. Payload 3.0: The first CMS that installs directly into any Next.js app. https://payload-cms.com/blog/payload-30-the-first-cms-that-installs-directly-into-any-nextjs-app, 2024.

Design eines performanten und deadlockfreien Sperrsystems für konkurrierende Zugriffe in einer objektorientierten Datenbank

Thomas Fetter **Harald Melcher**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma ctrl-s, Stuttgart

Einleitung

Die Firma ctrl-s GmbH hat mit symphony ein leistungsstarkes Enterprise Resource Planning (ERP) System entwickelt, das besonders für Druckkunden optimiert ist.

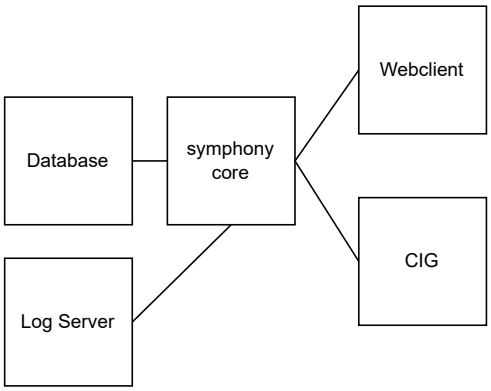


Abb. 1: Vereinfachte Darstellung der Symphony Komponenten [1]

Abbildung 1 zeigt die wesentlichen Komponenten von symphony. Die Echtzeitverwaltung erfolgen im Arbeitsspeicher des Kerns (symphony core), während der Datenbankserver als reines Backup dient. Webclients und der Cloud Integration Gateway (CIG) bilden die Schnittstellen zur Außenwelt. Das Logging ist, wie das Backup, aus Performancegründen auf separate Server ausgelagert, um weniger Ressourcen im Kern zu verbrauchen. symphonie-Objekte können beliebig komplex sein und zusätzlich miteinander verknüpft werden. In einem Webshop, in dem bedruckte Kaffeetassen mit eigenen Fotos bestellt werden können, stellt jede Tasse ein Objekt dar – ebenso der Drucker, der die Tasse bedruckt, oder die gesamte Bestellung. Workflows in symphony beschreiben den Übergang von einem Eingangszustand in einen Ausgangszustand. Sie können mehrere Pfade mit unterschiedlichen Zwischenzuständen enthalten. Ein Workflow könnte beispielsweise mit dem Zustand *Bestellungseingang* starten. Danach folgen Aktionen, die zu Folgezuständen führen, bis hin zum Zustand *an Kunde Zugestellt*. Bei jeder Zustandsänderung erfolgen zahlreiche Lese- und Schreibvorgänge auf den beteiligten symphony-Objekten. Um die Datenintegrität in symphony zu gewährleisten, ist eine geordnete Steuerung von Lese- und Schreibvorgängen entscheidend. Dies wird besonders wichtig, wenn Objekte gleichzeitig in mehreren Workflows verwendet werden können. Ein ungeordneter Zugriff würde zu Inkonsistenzen und Fehlern führen. Gleichzeitig muss sichergestellt werden, dass so viele Lese- und Schreibvorgänge wie möglich parallel ablaufen können, um die hohe Performanceanforderung 300.000 Objekt-Operationen pro Sekunde zu erfüllen [2]. Diese Bachelor Arbeit beschäftigt sich daher mit der Entwicklung eines Mechanismus zur Zugriffssteuerung, der Konflikte verhindert, eine hohe Parallelität ermöglicht und dabei maximale Effizienz sicherstellt.

Grundlagen

Das aktuell in symphony implementierte Locking-System ist auf einfache Schreibsperren für einzelne Objekte beschränkt (single-entity-centric write locks only) und weist zwei grundlegende Einschränkungen auf.

Erstens fehlt es an Funktionalität. Wichtige Mechanismen wie Ausführungssperren (Execute-Locks) sind nicht vorhanden. Diese sind jedoch essenziell, um eine effiziente Auslastungssteuerung zu ermöglichen.

Zweitens besteht ein erhöhtes Deadlock-Risiko: Da nicht mehrere Objekte gleichzeitig, sondern nur nacheinander gesperrt werden können, steigt die Wahrscheinlichkeit von Deadlocks erheblich [2]. Der ge-

nerelle Umgang mit nicht unterbrechbaren Ressourcen folgt dabei einem klaren Schema:

1. Ressource anfordern
2. Ressource benutzen
3. Ressource freigeben

Diese Abfolge stellt die Grundlage für jegliche Zugriffssteuerung dar. Es wird empfohlen, die Zeit zwischen dem Anfordern (1.) und dem Freigeben (3.) so kurz wie möglich zu halten, um die Effizienz zu maximieren und Konflikte zu vermeiden. [5] (S. 534 - 535)
Das neue Locking-System unterstützt drei primäre Sperrtypen:

1. Lese-Sperren (**Read-Locks**): Ermöglichen den gleichzeitigen Lesezugriff durch mehrere Tasks auf ein Objekt. Lese-Sperren blockieren Schreibzugriffe, erlauben jedoch weitere parallele Lesezugriffe.
2. Schreib-Sperren (**Write-Locks**): Blockieren sowohl Lese- als auch weitere Schreibzugriffe anderer Tasks. Schreib-Sperren gewährleisten Exklusivität für Änderungen.
3. Ausführungs-Sperren (**Execute-Locks**): Definieren eine maximale Anzahl von gleichzeitigen Zugriffen jeglicher Art auf eine Gruppe von Objekten (**Lock-Handle**). Sie fungieren ähnlich wie Semaphoren ([5] S. 177 - 179) und geben eine obere Schranke an gleichzeitig ausführbaren Lock-Handles an. Anders als Read- und Write-Locks beziehen sich Execute-Locks nicht auf einzelne Objekte, sondern dienen als übergeordnete Steuerungsebene, die parallele Zugriffe auf Gruppen von Objekten reguliert.

Darüber hinaus benötigen alle drei Sperren:

- Zeitliche Beschränkungen für Sperren: Jede Sperre verfügt über zwei zeitliche Parameter: Gather Time: Die maximale Dauer, innerhalb derer eine Lock-Anfrage erfolgreich sein muss. Nach Ablauf dieser Zeit verfällt die Anfrage automatisch. Maximum Lock Time: Die maximale Dauer, für die eine gewährte Sperre aktiv bleibt. Nach Ablauf dieser Zeit wird die Sperre automatisch freigegeben.
- Benachrichtigung bei Ablauf der Sperre: Nach 80% der Maximum Lock Time muss das System eine Warnung ausgeben, dass die Sperre bald abläuft.
- Effizientes Management von Sperrgruppen: Während ein Lock-Handle gesperrt ist, müssen individuelle Sperranfragen auf einzelne Objekte innerhalb dieser Gruppe gezielt entsperrt werden können.

Implementierung

Ein **Lock-Request** ist eine Zugriffsanfrage auf ein individuelles symphony-Objekt. **Lock-Handles** hingegen repräsentieren eine Liste solcher Anfragen und ermöglichen so das Verwalten mehrere Lock-Requests in einer Operation. Dies verhindert die Gefahr von Deadlocks und erhöht zugleich die Performance, da Zugriffsrechte für mehrere Ressourcen gleichzeitig vergeben werden können, anstatt einzeln für jede Ressource. Die Beziehung der gewählten Daten-Typen ist in Abbildung 2 zu sehen.

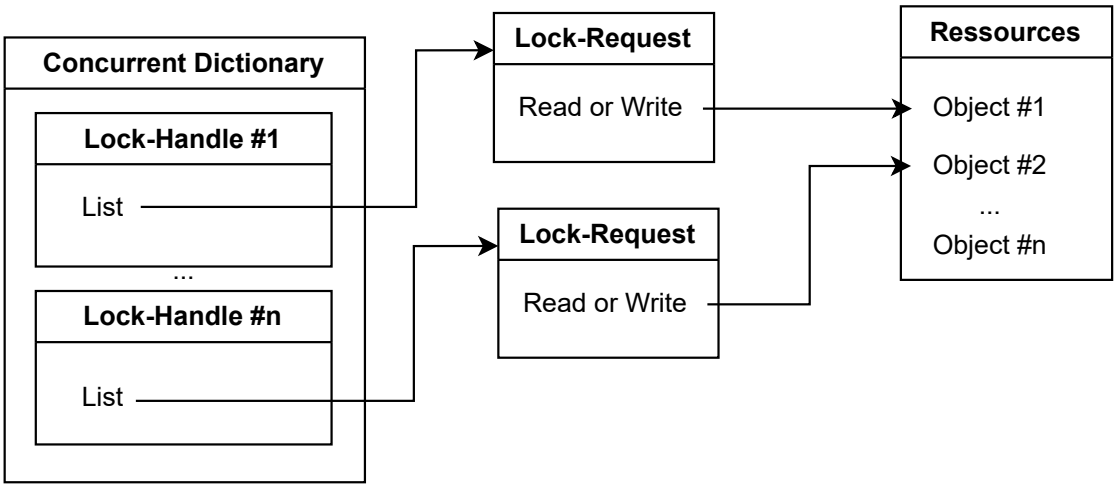


Abb. 2: Datenstruktur zum Verwalten von Zugriffsanfragen [1]

Lock-Handles werden nicht recycelt und sind einmalig verwendbar.
Der Lebenszyklus eines Lock-Handles wird im Lock-State definiert. Dieser Status gilt einheitlich für alle Lock-Requests eines Lock-Handles:

- WaitForLock – Die Zugriffsanfrage wurde gestellt und wartet auf Freigabe.
- Locked – Der Zugriff wurde erteilt.
- Released – Die Daten werden vom Prozess nicht mehr benötigt und wurden freigegeben.
- Cancelled – Die Zugriffsanfrage wurde vom Server oder Client abgebrochen.

Zur Verwaltung der Sperren werden threadsichere Datenstrukturen wie Concurrent-Dictionary [3] verwendet (auch in Abbildung 2 zu sehen), die in C# durch .NET bereitgestellt werden. Diese ermöglichen eine sichere und effiziente Verwaltung der Status von Lock-Handles, selbst wenn mehrere Threads parallel auf sie zugreifen. Es sei erwähnt, dass nur das Hinzufügen und Entfernen von Elementen in dem Concurrent-Dictionary von Haus aus Threadsicher sind [3]. Für das Verändern von einzelnen Attributen eines Dictionary Values muss die Threadsicherheit vom Algorithmus gewährleistet werden.
Die Lock-Handles verfügen über eine asynchrone Wait()-Methode. Mithilfe einer TaskCompletionSource [4] wird ein Token erstellt, das den wartenden Task informiert, sobald alle angeforderten Lock-Requests verfügbar sind. Dadurch wird sichergestellt, dass Tasks effizient und nicht-blockierend auf die Freigabe der benötigten Ressourcen warten können.

Ausblick

Mit den aktuellen zusätzlichen Funktionen erfüllt das Locking-System bereits die Funktionalität der Execute-Locks und die Möglichkeit, mehrere Sperren (Lock-Requests) eines Lock-Handles in einer einzigen Operation zu setzen. Dennoch gibt es Potenzial für Optimierungen und Erweiterungen:

- Optimierung der Warteschlange: Die Einführung einer Prioritätswarteschlange könnte kritische Anfragen bevorzugt behandeln und die Systemeffizienz weiter steigern.
- Parallele Verarbeitung unabhängiger Anfragen: Aktuell wird die Warteschlange sequentiell bearbeitet. Zukünftige Implementierungen könnten parallele Zugriffe auf unabhängige Lock-Handles erlauben.
- Alternative zum Execute-Lock: Eine genauere Analyse könnte zeigen, ob die Aufgaben des Execute-Locks effizienter in eine separate Scheduling-Schicht ausgelagert werden könnten.

Sollte sich bei Performancetests herausstellen, dass die Verwendung von Concurrent-Dictionaries nicht den Anforderungen entspricht, müssen spezialisierte, threadsichere Datenstrukturen entwickelt werden, um die Effizienz und Skalierbarkeit des Systems zu gewährleisten.
Das entwickelte Locking-System bildet eine solide Grundlage für die effiziente und zuverlässige Ressourcenverwaltung in symphony. Es erfüllt zentrale Anforderungen wie Execute-Locks und Deadlock-Vermeidung und bietet Potenzial für zukünftige Optimierungen, um wachsenden Performanceansprüchen gerecht werden zu können.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Matthias Lukaseder. Symphony Core 1.7 locking mechanism. https://ctrl-s.atlassian.net/wiki/spaces/SYMPHONYCORE/pages/4089380886/Symphony+Core+1.7+locking+mechanism, 04 2024.

[3] Dokumentation Microsoft. ConcurrentDictionary<TKey,TValue> Klasse. https://learn.microsoft.com/de-de/dotnet/api/system.collections.concurrent.concurrentdictionary-2?view=net-8.0, 11 2024.

[4] Dokumentation Microsoft. TaskCompletionSource<TResult> Klasse. https://learn.microsoft.com/de-de/dotnet/api/system.threading.tasks.taskcompletionsource-1?view=net-8.0, 11 2024.

[5] Andrew S. Tanenbaum and Herbert Bos. *Moderne Betriebssysteme*. Pearson Education, 4 edition, 2016.

Evaluierung eines BPMN-Low-Code-Plattformprototyps zur Prozessentwicklung einer Smart-Factory hinsichtlich der Senkung des Kompetenzbedarfs

Marcel Fetzer **Dieter Morgenroth**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Adesso SE, Stuttgart

Motivation

Durch die Einführung der Industrie 4.0, eröffnet sich für Unternehmen die Möglichkeit, neue Technologien zur Digitalisierung und Vernetzung der am Wertschöpfungsprozess beteiligten Akteure zu implementieren. Durch die digitale Vernetzung der einzelnen Akteure, können die Wertschöpfungsprozesse besser aufeinander abgestimmt werden und somit eine Effizienzsteigerung der gesamten Wertschöpfung des Unternehmens erreicht werden [7]. Trotz dessen, dass die Integration der Industrie 4.0, für viel Unternehmen in einer verbesserten Wettbewerbsfähigkeit resultiert, kann diese zugleich eine hohe Belastung für das jeweilige Unternehmen bedeuten. Hierbei spielen unter anderem fehlende finanzielle Mittel, fehlende Fachkräfte sowie die Komplexität des Themas eine große Rolle [5]. Um die innerhalb des Unternehmens ablaufenden Prozesse für eine breitere Masse an Personen verständlich zu machen, kann auf standardisierte Notationsformen wie beispielsweise die BPMN zurückgegriffen werden. Durch die Verwendung dieser, können beispielweise die Prozesse einer Smart-Factory, vereinheitlicht abgebildet werden. Dies erhöht die Verständlichkeit der Prozesse im Unternehmen und reduziert somit die Komplexität deren Umsetzung [3]. Ergänzend dazu, kann eine Kombination des Low/No-Code Ansatzes zum Entwurf einer Entwicklungsplattformen eingesetzt werden. Hierdurch würde die Entwicklung von Programmcode, durch die logische Kombination vordefinierter Funktionsblöcke sowie der Einbindung kurzer Code-Fragmente ermöglicht. Die mittels dieses Ansatzes erreichte Abstraktionsebene, gegenüber dem im Hintergrund implementierten Quellcode, kann für eine Komplexitätsreduktion der Entwicklung von Softwareanwendungen sorgen [8]. Eine Kombination der beschriebenen Ansätze, in Form einer Modellierungsplattform, könnte bei der Lösung der genannten Probleme von großer Hilfe sein.

Stand der Forschung

Der Einsatz von Low-Code und die damit einhergehenden Vorteile zur Unterstützung der digitalen Transformation von Unternehmensprozessen, konnten wie [1] zeigt, bereits in einzelnen Studien untersucht und nachgewiesen werden. In Kombination dazu, kann die BPMN wie in [6] beschrieben, durch ihre Standardelemente, die IoT-Fähigkeit von Systemen in BPMN-Prozessmodelle integrieren. Da die Abbildung einfacher IoT-Funktionalität schnell zu komplexen Prozessmodellen führen kann, gibt es Ansätze die BPMN zur Abbildung komplexerer Prozesse, wie beispielsweise eines Cyber-Physischen Systems zu erweitern [4]. Hier soll der im Rahmen der Thesis entwickelte Plattform-Prototyp ansetzen und so die Abstraktionsebene, unter dem Ziel der Komplexitätsreduzierung maximieren. Der Fokus der Thesis liegt hierbei auf der Bewertung des Prototyps, unter den Gesichtspunkten einer Low/No-Code Plattform.

Zielsetzung

Das Ziel der Thesis liegt in der Beantwortung der Frage, ob eine auf der BPMN aufbauende Low/No-Code Anwendung zur Modellierung eines Smart-Factory Prozesses die zuvor erwähnten Belastungen der Unternehmen zur Einführung der Industrie 4.0 reduzieren kann. Hierbei wird der Fokus, auf die eingangs angesprochenen Problemstellen der Systemkomplexität sowie des Fachkräftemangels gelegt. Um eine Antwort auf diese Frage zu finden, wird eine auf der BPMN basierende Low/No-Code Plattform entwickelt, mit deren Hilfe der Prozess einer Smart-Factory modelliert und anschließend auf dieser ausgeführt werden kann. Die Plattform bildet einen Proof-of-Concept und wurde zur Erreichung des verfolgten Zieles entwickelt. Um die Abstraktionsebene zu erhöhen und die Plattform in ihrer Komplexität zu reduzieren, wird die herkömmliche BPMN, um neue Elemente erweitert. Die hierzu

verwendete Logik soll wie in Abbildung 1 dargestellt, ein unter Verwendung der BPMN-Erweiterungselementen im Frontend modelliertes Prozessmodell, über einen Backend-Service, in die den Elementen entsprechenden Zustandsübergänge konvertieren, um daraus eine Zustandsmaschine zu initialisieren. Die aus der Konvertierung als Output resultierende Zustandsmaschine, übernimmt die Aufgabe der Orchestrierung des Fabrikprozesses.

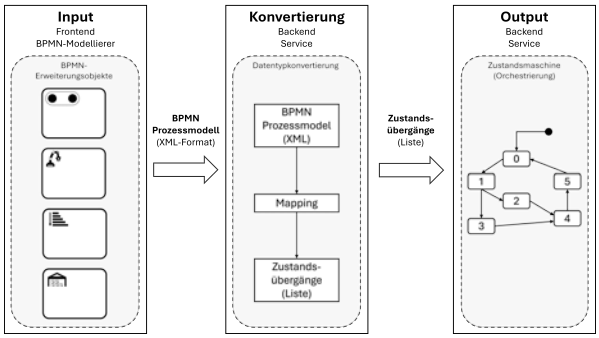


Abb. 1: Konvertierung des Prozessmodels in einen neuen Datentyp [2]

Aufbau der Methodik

Die zur Modellierung und Programmierung des Produktionsprozesses entwickelte Low/No-Code Plattform, soll durch die Nutzer hinsichtlich des definierten Ziels, der Komplexitätsreduzierung zur Programmierung eines Prozessablaufes innerhalb einer Smart-Factory bewertet werden. Die hierfür benötigten Bewertungsergebnisse, werden mittels der Methodik des Fragebogens erhoben. Um die aufgestellte Forschungsfrage beantworten zu können, werden die Ergebnisse ausgewertet und mit der aufgestellten These verglichen. Die zur Bewertung der Anwendung befragten Personen bekommen hierzu eine oberflächliche Erklärung der Modellierungsplattform sowie des Aufbaus der physischen Anlage. Zur Bewertung könnten unter anderem Faktoren, wie die benötigte Bearbeitungszeit der gestellten Aufgaben oder der Intuitionsgrad des Aufbaus der Anwendung gemessen werden. Um aussagekräftige Daten zu erhalten, sollten Personen mit unterschiedlichem Verständnis über die im Hintergrund ausgeführten Logik als Teilnehmer der Umfrage herangezogen werden. Die zur Erhebung der Nutzungserfahrung entwickelte Umfrage, besteht im Kern aus der Bewertung, der wie in Abbildung 2 dargestellten Oberfläche des Prototyps. Die dargestellte Oberfläche soll dem Nutzer über unterschiedliche Bereiche einen intuitiven Prozessentwurf ermöglichen. Hierzu gehört die Bereitstellung der Modellierungsobjekt über die Objektpallette, die Modellierung der Prozessmodelle auf dem Modellierungs-Canvas, die spezifische Anpassung von Elementeigenschaften über das Eigenschaftspanel sowie das Speichern, Exportieren oder Ausführen über den Ausführungsbereich.

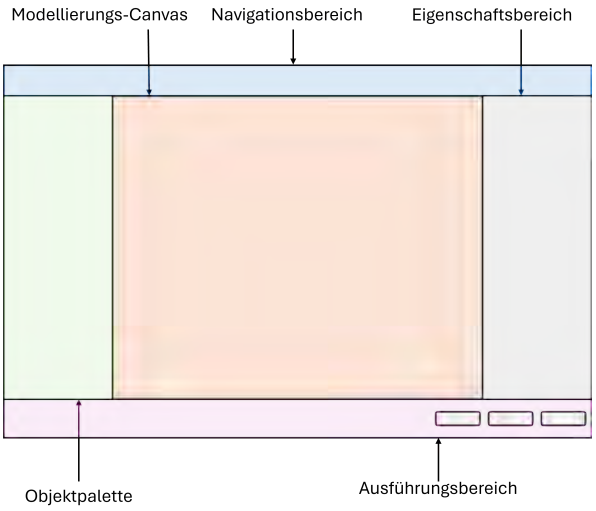


Abb. 2: Aufbau des Modellierungstool [2]

Ausblick

Da zum aktuellen Zeitpunkt noch keine spezifischen Umfrageergebnisse zur Funktionsweise der Plattform vorliegen, kann nur ein theoretischer Ausblick gegeben werden. Dieser Ausblick könnten demnach wie folgt aussehen. Nach Auswertung der Umfragedaten, könnten weitergehende Schritte geplant sowie aus den Bewertungen resultierende Anforderung erarbeitet werden. Im Anschluss an die Arbeit könnte der Proof-of-Concept weiter ausgebaut und durch neue Funktionalität erweitert werden. Mögliche Funktionalitäten könnten die vereinfachte Integration neuer Maschinen in das Prozessmodellierung umfassen.

Literatur und Abbildungen

[1] Z. Cai et al. A Case Study: Digitalization of Business Processes of SMEs with Low-Code Method. In *IFAC-PapersOnLine*, volume 55, pages 1840–1845. Elsevier, 2022.

[2] Eigene Darstellung.

[3] Jakob Freund and Bernd Rücker. *Praxishandbuch BPMN 2.0*. Hanser, 3 edition, 2012.

[4] Imen Graja et al. BPMN4CPS: A BPMN Extension for Modeling Cyber-Physical Systems. In *2016 IEEE 25th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*, pages 152–157. IEEE, 2016.

[5] Angelina Marko. Industrie 4.0 - so digital sind Deutschlands Fabriken. https://www.bitkom.org/sites/main/files/2023-01/221125StudieIndustrie-40-1.pdf, 09 2022.

[6] Francisco Martins and Dulce Domingos. Modelling IoT behaviour within BPMN Business Processes. In *Procedia Computer Science*, volume 121, pages 1014–1022. Elsevier B.V, 2017.

[7] Armin Roth. Industrie 4.0 – Hype oder Revolution? In *Einführung und Umsetzung von Industrie 4.0: Grundlagen, Vorgehensmodell und Use Cases aus der Praxis*, pages 1–15. Springer Gabler Berlin, Heidelberg, 1 edition, 2016.

[8] Stefan Sauer, Nils Weidmann, and Jonas Kirchhoff. Merkmale und Entwicklungslinien der Low-Code-Programmierung. In *Prozesse in Industriebetrieben mittels Low-Code-Software digitalisieren: Ein Praxisleitfaden*, pages 17–29. Springer Vieweg Berlin, Heidelberg, 1 edition, 2023.

An Analysis of High-Resolution Feature Maps for Monocular Depth Estimation

Florian Fink **Markus Enzweiler**

Department of Computer Science and Engineering, Esslingen University

Work carried out at Mercedes-Benz AG, Sindelfingen

Introduction

In recent years, there has been a notable increase in the utilisation of machine learning within the industrial sector. In the context of autonomous driving, machine learning algorithms are frequently used for the purpose of identifying entities such as road users and traffic signs, as well as route planning, map acquisition, and obstacle distance estimation. Monocular depth estimation represents a practical application of machine learning, whereby the objective is to estimate the distance of each pixel in an image. In contrast to stereo depth estimation, monocular depth estimation utilises a single image to determine distances. Figure 1 illustrates two example images and the corresponding depth predicted with such a model. Monocular depth estimation is employed in a variety of contexts, including social media applications and robotics, as well as in autonomous driving. The estimation of distances is a fundamental task in this context, as the vehicle requires such information to make driving decisions and therefore minimise the risk of collisions. While it may appear to be a simple task for the human brain, it is a challenging problem for computers, especially with just one camera perspective. In order to complete this task, the machine must learn a sophisticated understanding of the given environment. This understanding can also be applied to other tasks that require an understanding of the surrounding context, such as object recognition.

Motivation

The primary challenges associated with monocular depth estimation can be broadly classified into three categories: the collection of suitable data, the ability to generalise across different datasets and the high computational cost, which in turn affects the real-time capability. To reduce the computational effort, neuronal networks typically downscale high-resolution images to low-resolution features. While this approach improves computation, it also results in a loss of sharpness. A network which scales up low-resolution features to a higher resolution, also known as a feature upsampler, could potentially enhance depth estimation by increasing sharpness with reduced computational cost.

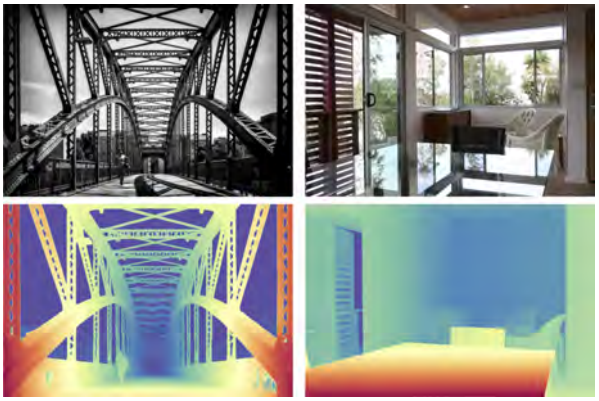


Fig. 1: Input Images and visualised predicted depths with Depth Anything V2 [3]

Research Objectives

The objective of this thesis is to investigate the potential of the state-of-the-art (SOTA) feature upsampler called FeatUp, especially for autonomous driving. In particular, the following research questions are posed:

1. Does FeatUp enhance the performance of state-of-the-art monocular depth estimation models?
2. What is the optimal method for integrating FeatUp into a state-of-the-art monocular depth estimation model?
3. Is it feasible to reduce the upscaling factor of FeatUp?
4. What is the effect of image size on the final result?

Background

Depth Anything V2 is a state-of-the-art monocular depth estimation model consisting of a DINOv2 Encoder and a DPT Decoder. While the training procedure of the model is more complex, the retraining and final usage are simple in comparison to other state-of-the-art models. Due to these characteristics, the implementation of the upsampler is more straightforward, while achieving SOTA results. [3]
The FeatUp feature upsampler, as proposed by the paper "FeatUp a Model-Agnostic Framework for Features at Any Resolution" comprises two versions: FeatUp (JBU), which is the more general approach and FeatUp (Implicit), which overfits on a specific image. The upsamplers are trained to enhance low-resolution features with details from the input image, and therefore producing high-resolution features. [1]

Implementation

In order to evaluate the performance of FeatUp, the input image is resized using the same factor by which FeatUp upsamples the features. As illustrated in Figure 2, the upscaled image is fed into the Depth Anything V2 Network, while the original image is fed into the Depth Anything V2 Network with FeatUp. Ideally, the output of both versions would be identical. The FeatUp version should have a significantly reduced computational cost due to the smaller image input size of the encoder. Upscaling the image does not change the image information, allowing for a direct comparison between the two methods.

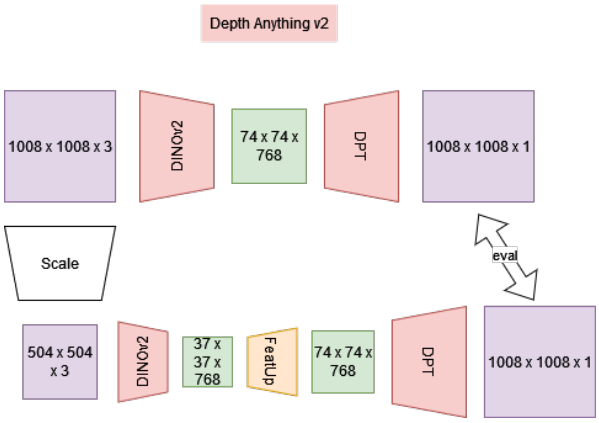


Fig. 2: Network structure used to evaluate the impact of FeatUp [2]

To evaluate and compare the various approaches, different evaluation metrics are employed. As details are an important aspect of the upsampler a specific edge sharpness metric is applied.
As Figure 2 is merely a simplified representation, the upsampler is implemented in a variety of ways and different parts of the network are trained to get disparate results.

Outlook

In order to assess the generalisability of the findings, future work could evaluate the impact of varying depth estimation models, datasets and upsamplers.

References and figures

[1] Stephanie Fu et al. FeatUp: A Model-Agnostic Framework for Features at Any Resolution. https://arxiv.org/abs/2403.10516, 2024.

[2] Own representation.

[3] Lihe Yang et al. Depth Anything V2. https://arxiv.org/abs/2406.09414, 2024.

Impact of Data Reduction on Model Performance in HD Map Datasets

Daniel Fritz **Steffen Schober**

Department of Computer Science and Engineering, Esslingen University

Work carried out at Mercedes-Benz AG, Böblingen

Introduction

A key aspect of autonomous driving is the continuous awareness of the vehicle's road environment. Digital maps commonly used in navigation systems are insufficient to meet the requirements for an accurate and reliable representation of the real-world road, particularly for deployment in autonomous vehicles (AVs). The utilization of sensors, such as radar, IMU, GPS, and camera enables the creation of a more accurate digital map, also known as a high-definition (HD) map. HD maps provide a precise representation of the environment, thereby enabling the next logical decision to be made in a variety of autonomous driving scenarios. Once an HD map has been created and made available to an AV, it is often considered a hidden or virtual sensor, as it uses knowledge from various sensors to build the vehicle's road environment [3]. An example of an HD map is depicted in Fig. 1.

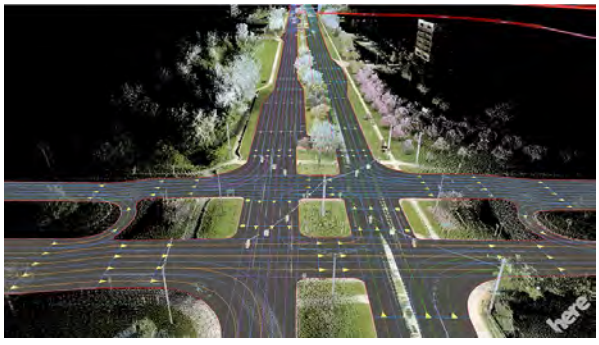


Fig. 1: Example of the HD map [1]

Motivation and Goal

HD maps are characterized by their highly accurate representations of road models. This requires high-frequency sampeling rates from sensors to obtain a sufficient number of data points representing the roads driven on. In addition, crowdsourced data is processed to keep the AV's map up to date. This creates a storage burden, as HD maps require a lot of data to sufficently describe the AV's road environment. Performing complex preprocessing and data analysis on large datasets can result in high computational costs, making such operations impractical or infeasible. Therefore, it is desirable to reduce the dataset in a certain manner that preserves essential information while minimizing data redundancy. The use of data reduction techniques facilitates subsequent preprocessing steps in terms of computational time and can potentially result in higher accuracy in the prediction of upcoming road segments by means of Transformers [9] or Graph Neural Networks (GNNs) [8]. The objectives of the thesis are the following:

- Identifying suitable data reduction methods to reduce the given dataset while maintaining its essential characteristics.
- Building a pipeline for dynamic exploration of various data reduction methods to obtain their impact and calculating their error metrics to evaluate the performance in terms of information loss and computation time.
- Training models based on Transformer and GNN, and comparing their results using the original dataset versus a reduced dataset.

Related Work

In their paper, Mink et al. [7] proposed Lane Model Transformer Network (LMT-Net) to derive lane graphs based on observations from vehicles on the road. These observations consist of different traces representing properties of the road, including road boundaries, dashed lines, and road centerlines. Each trace is represented as a polyline, which is described by successive spatial points. Multiple polylines, which can ultimately be seen as a graph, are used as input to the Transformer to derive a road model. Subsequently,

a comparison is made between predicted roads and manually annotated ground truth roads.

Data Reduction

Data reduction includes different categories; one of them is data compression. Data compression methods use algorithms to remove samples of the dataset. If the dataset can be reconstructed after the compression process, then the data reduction is called lossless. If, instead, the dataset can only be approximately reconstructed, then the data reduction is called lossy. Lossless algorithms allow only limited data manipulation while ensuring the complete reconstruction of the original data. In contrast, lossy algorithms allow more flexibility and can be tailored to specific tasks, deciding how aggressively data reduction should be applied [5]. A well-known lossy algorithm for data compression is the Douglas-Peucker (DP) [2] algorithm. DP aims to generalize trajectories (polylines) by removing less informative points. The result is a simplified trajectory with fewer points while remaining the essential characteristics of the original one. Several papers have been published on how the basis of the DP algorithm was extended to achieve better results in terms of the removal ratio of points, the information loss, and the time complexity.
DP provides promising results on the task of trajectory simplification, but it is not capable of aggregating multiple trajectories, which contributes significantly to data redundancy. This issue can be addressed by the algorithm TRACLUS [6]. The main idea behind TRACLUS is to create a representative trajectory based on multiple trajectories. This is archived by first creating clusters of similar trajectories and then aggregating each cluster into a single trajectory, as depicted in Fig. 2. Since trajectories can have large dissimilarities, i.e., two trajectories moving away from each other, clustering over a global scope misses similarities at a segment level. Therefore, TRACLUS splits the trajectories into segments and performs clustering at the segment level, thus addressing local similarities. Thereby, the clustering algorithm is based on DBSCAN, adapted to operate with line segments rather than points.

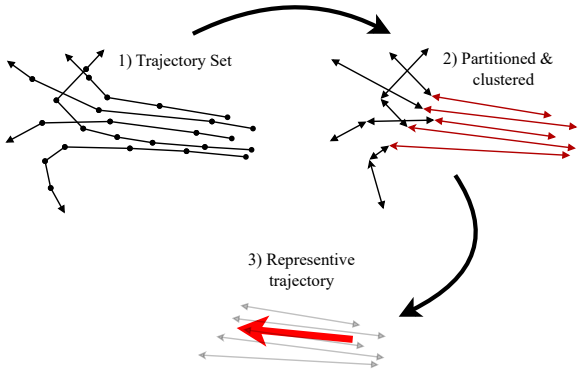


Fig. 2: An example of partitioning and clustering trajectories [6]

Constructing Lane Graphs

Inferring a lane graph requires processing the given dataset to ensure its compatibility with the neural network's expected input format. Since a GNN expects graphs as input, the data must be tailored accordingly. As previously mentioned, a polyline consists of consecutive spatial points, which can be converted into a graphical representation. Formally, a graph can be denoted as $G = (V, E)$, where V is a set of nodes and E is a set of edges describing the connection between nodes. An existing connection between the nodes $u \in V$ and $v \in V$ is denoted as $(u, v) \in E$ [4].
Applying these definitions to the dataset, a graph can be constructed from the given polylines. Thus, a GNN can be utilized to transfer a graph consisting of multiple polylines to a latent representation. Subsequently, based on the latent representation, a corrected graph can be constructed and compared with the ground truth graph.

Outlook

Architectures such as Transformers and GNNs can be employed to predict a lane model based on the obtained polylines. Training with the reduced dataset should be conducted, and the impact of data reduction on both accuracy and training time must be analyzed. Furthermore, hyperparameter optimization should be performed on both data reduction and training in order to achieve optimal results with the least amount of data.

References and figures

[1] Pino Bonetti. HERE introduces HD Live Map to show the path to highly automated driving. https://www.here.com/learn/blog/here-introduces-hd-live-map-to-show-the-path-to-highly-automated-driving, 01 2016.

[2] David Douglas and Thomas K. Peucker. Algorithms for the Reduction of the Number of Points Required to Represent a Digitized Line or Its Caricature. *Cartographica*, pages 112–122, 1973.

[3] Gamal Elghazaly, Raphaël Frank, Scott Harvey, and Stefan Safko. High-Definition Maps: Comprehensive Survey, Challenges, and Future Perspectives. *IEEE Open Journal of Intelligent Transportation Systems*, pages 527–550, 2023.

[4] William L. Hamilton. *Graph Representation Learning*. Springer International Publishing, 2022.

[5] Jiawei Han and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 3 edition, 2012.

[6] Jae-Gil Lee and Jiawei Han. Trajectory clustering: a partition-and-group framework. *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pages 593–604, 2007.

[7] Michael Mink, Thomas Monninger, and Steffen Staab. LMT-Net: Lane Model Transformer Network for Automated HD Mapping from Sparse Vehicle Observations. *arXiv*, 2024.

[8] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The Graph Neural Network Model. *IEEE Transactions on Neural Networks*, pages 61–80, 2009.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *arXiv*, 2023.

Identifikation eines optimierten KI-Algorithmus zur Fehlererkennung in industriellen Bildern mit geringer NIO-Bilderanzahl

Ismet Gezer **Mirko Sonntag**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einführung

Die industrielle Bildverarbeitung stellt einen wesentlichen Bestandteil der Qualitätssicherung in modernen Fertigungsprozessen dar. In Branchen, in denen hohe Präzisionsanforderungen bestehen, wie beispielsweise in der Automobilindustrie oder der Elektronikindustrie, ist der Einsatz automatisierter Fehlererkennungssysteme unerlässlich, um Ausfälle im Produktionsprozess zu minimieren und somit die damit verbundenen Kosten zu reduzieren. Der Einsatz künstlicher Intelligenz (KI) erlaubt die Automatisierung komplexer visueller Prüfungen und eine signifikante Beschleunigung von Entscheidungsprozessen. [9]
Eine wesentliche Herausforderung besteht in der limitierten Verfügbarkeit von Bildern, die Produktionsfehler aufweisen (NIO). Während fehlerfreie Bilder (IO) in der Regel in großer Zahl verfügbar sind, sind NIO-Bilder aufgrund seltener Defekte nur begrenzt zugänglich. Dies erschwert den Trainingsprozess gängiger KI-Modelle, die auf großen, ausgewogenen Datensätzen basieren. [2] Ein zunehmend an Bedeutung gewinnender Ansatz ist die Generierung synthetischer NIO-Bilder zur Erweiterung der begrenzten Datenbasis. In der vorliegenden Arbeit wird untersucht, inwiefern sich KI-Modelle durch reale und synthetische NIO-Bilder robust trainieren lassen, um trotz unbalancierter Datensätze präzise Ergebnisse zu erzielen.

Problemstellung

Industrielle Bildverarbeitungssysteme sehen sich häufig mit der Herausforderung konfrontiert, dass die Datensätze eine hohe Unausgewogenheit aufweisen. Dabei ist die Anzahl der NIO-Bilder im Vergleich zu den IO-Bildern sehr gering. Die Disparität führt dazu, dass KI-Modelle dazu neigen, die Mehrheitsklasse zu bevorzugen und Minderheitsklassen zu vernachlässigen, was in der Praxis zu hohen Fehlerraten führen kann. Dies ist insbesondere in sicherheitskritischen Anwendungen inakzeptabel.
Des Weiteren stellt die Generierung synthetischer NIO-Bilder eine technisch anspruchsvolle Aufgabe dar. Der Erfolg solcher Bilder ist maßgeblich von ihrer realistischen Darstellung sowie ihrem semantischen Informationsgehalt abhängig. Die vorliegende Arbeit befasst sich daher mit der Untersuchung der Einbindung synthetischer NIO-Bilder in den Trainings- und Evaluierungsprozess, mit dem Ziel, die Robustheit und Genauigkeit der Modelle zu erhöhen.

Methodische Ansätze

Die Entwicklung eines robusten KI-Algorithmus zur Fehlererkennung bedingt den Einsatz zeitgemäßer maschineller Lernverfahren, welche sowohl überwachte als auch unüberwachte Ansätze umfassen. Die Voraussetzung für überwachtes Lernen ist die Verfügbarkeit einer hinlänglichen Menge gelabelter Trainingsdaten. Convolutional Neural Networks (CNNs) erweisen sich aufgrund ihrer Fähigkeit, visuelle Merkmale hierarchisch zu erlernen, als die bevorzugte Wahl in der industriellen Bildverarbeitung [4]. Abbildung 1 zeigt die grundlegende Architektur eines CNNs, bestehend aus Schichten für Faltung, Aktivierung und Pooling. Diese Netzwerke extrahieren zunehmend komplexere Merkmale aus Eingabebildern und ermöglichen so eine präzise Klassifikation industrieller Bilddaten. Architekturvarianten wie ResNet und EfficientNet sind durch hohe Präzision und Effizienz gekennzeichnet und haben sich als bewährte Methoden zur visuellen Fehlererkennung etabliert. [10] Eine wesentliche Herausforderung besteht jedoch darin, dass für das überwachte Lernen eine signifikante Anzahl korrekt gelabelter NIO-Bilder benötigt wird, was in der Praxis selten gegeben ist.

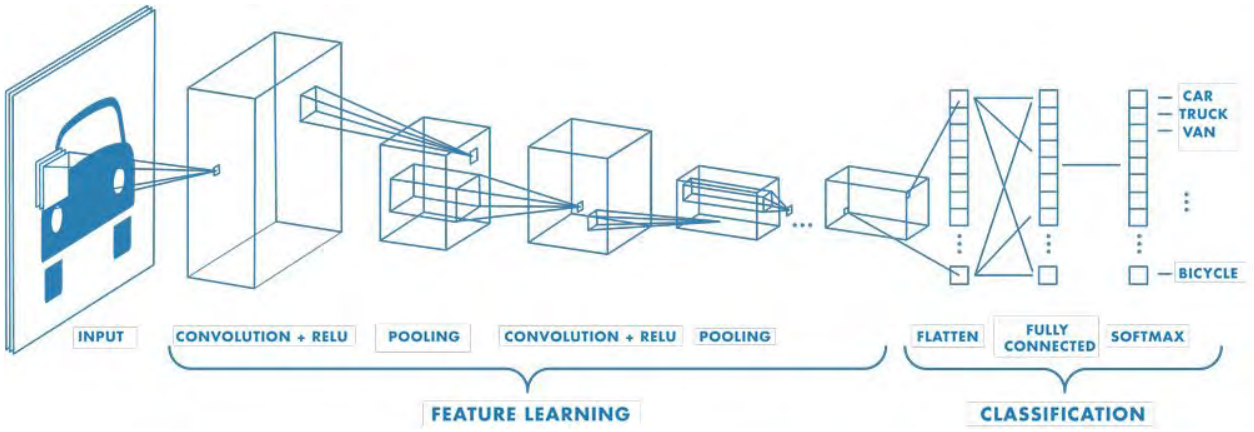


Abb. 1: Architektur - Convolutional Neural Network [6]

Unüberwachte Lernverfahren stellen eine alternative Herangehensweise dar, insbesondere wenn die Verfügbarkeit gelabelter Daten begrenzt ist. Autoencoder und Variational Autoencoders (VAEs) erlernen latente Merkmalsdarstellungen, indem sie Anomalien durch Abweichungen vom gelernten Normalzustand erkennen. [3] Abbildung 2 zeigt die typische Struktur eines Autoencoders mit Encoder- und Decoder-Netzwerken, die Bilder komprimieren und rekonstruieren. Abweichungen zwischen Eingabe- und Ausgabebildern weisen dabei auf potenzielle Fehler hin. Clustering-Methoden wie k-Means oder DBSCAN ermöglichen die Gruppierung von Bildern basierend auf visuellen Ähnlichkeiten und die Identifikation potenziell fehlerhafter Bilder ohne manuelle Annotation. [1] In der industriellen Praxis können unüberwachte Modelle als Vorfilter zur Fehlererkennung eingesetzt werden.

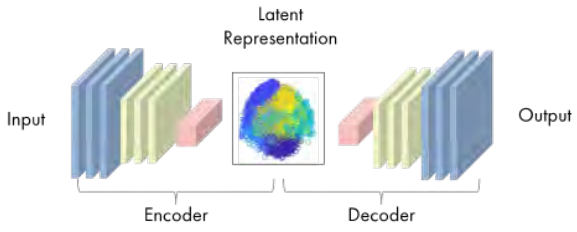


Abb. 2: Architektur - Autoencoder [5]

Ein wesentlicher Aspekt der Methodik ist die Integration synthetischer NIO-Bilder. Diese Integration erweitert die Datengrundlage und erlaubt die Gestaltung überwachter Modelle mit höherer Robustheit, indem selten auftretende Fehler explizit simuliert werden. [8] Im Rahmen des überwachten Lernens fungieren synthetische NIO-Bilder als zusätzliche Trainingsbeispiele, sodass der Algorithmus auch bei stark unbalancierten Datensätzen präzise Fehlererkennungen durchführen kann.
Im Rahmen des unüberwachten Lernens erfolgt keine direkte Verwendung von synthetischen NIO-Bildern für das Training, da das Modell auf Basis fehlerfreier Bilder den Normalzustand erlernt. Dennoch spielen synthetische NIO-Bilder eine zentrale Rolle in der Evaluierung, da sie zur Validierung der Anomalieerkennungsleistung eingesetzt werden können. Dies ermöglicht eine gezielte Überprüfung der Modellleistung anhand realitätsnaher Simulationsdaten. Die Bewertung erfolgt anhand etablierter Metriken wie Präzision, Recall und F1-Score, welche sowohl die Fähigkeit zur Fehlererkennung als auch die Minimierung falscher Alarme berücksichtigen. [7]

Ziel der Arbeit

Die vorliegende Arbeit verfolgt das Ziel, einen KI-Algorithmus zu entwickeln, der in der Lage ist, Defekte in industriellen Bildverarbeitungsanwendungen auch bei begrenzter Verfügbarkeit fehlerhafter Bilder zuverlässig zu erkennen. Der Schwerpunkt liegt auf der Optimierung bestehender Modellarchitekturen sowie der Erprobung spezifischer Verfahren zur Datenvorbereitung. Insbesondere wird der Einfluss synthetisch generierter NIO-Bilder untersucht, um die Modellleistung zu verbessern. Das Ziel besteht in der Kombination realer und künstlicher Daten, um eine präzise Fehlererkennung sicherzustellen und die Robustheit des Systems auch in stark unbalancierten Datensätzen zu gewährleisten.

Literatur und Abbildungen

[1] Charu C Aggarwal and Chandan K. Reddy. *Data clustering: algorithms and applications*. Chapman and Hall/CRC, 2014.

[2] Ian Goodfellow, Aaron Courville, and Yosuha Bengio. *Deep learning*. The MIT Press, 2016.

[3] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. *arXiv preprint*, 2013.

[4] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 2015.

[5] MathWorks. Was ist ein Autoencoder? https://de.mathworks.com/discovery/autoencoder.html, 2024.

[6] MathWorks. Was ist ein Convolutional Neural Network? https://de.mathworks.com/discovery/convolutional-neural-network.html, 2024.

[7] David M. W. Powers. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2020.

[8] Connor Shorten and Taghi M. Khoshgoftaar. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 2019.

[9] Richard Szeliski. *Computer Vision: Algorithms and Applications*. Springer International Publishing, 2022.

[10] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *International Conference on Machine Learning*, 2019.

Fehlerdichte als Metrik für Softwarequalität

Luisa Glass **Dirk Hesse**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Vector Informatik GmbH, Stuttgart

Einleitung

Mit der wachsenden Bedeutung von Software in sämtlichen Lebensbereichen wird die Sicherstellung von Softwarequalität immer zentraler. Softwarefehler, die zu Sicherheitsrisiken oder wirtschaftlichen Schäden führen, sind dabei von besonderer Relevanz. Um die Qualität von Software objektiv bewerten und quantifizieren zu können, kommen verschiedene Metriken zum Einsatz, darunter die Fehlerdichte [1]. Diese misst das Verhältnis von Anzahl Fehlern zur Größe der Software [3]. Die Größe der Software wird dabei häufig in *Lines of Code (LOC)* bestimmt. Dabei werden die in der Software gefundenen Fehler über einen Beobachtungszeitraum aufsummiert.
Trotz der verbreiteten Fokussierung auf *LOC* wird die Fehlerdichte auch durch weitere Faktoren wie Entwicklungszeitpunkt, Beobachtungszeitraum und Branche beeinflusst. Ziel ist es daher, den Einfluss dieser Faktoren auf die Fehlerdichte zu untersuchen.
Die zentrale Forschungsfrage dieser Arbeit lautet: „Wie wird die Fehlerdichte als Metrik zur Bewertung der Softwarequalität durch unterschiedliche Kontexte wie Entwicklungszeitpunkt, Beobachtungszeitraum, Größe der Software und Branche beeinflusst?“
Es wurden folgende Hypothesen aufgestellt:

- Je jünger der Entwicklungszeitpunkt der Software ist, desto geringer ist die Fehlerdichte
- Die Fehlerdichte ist unabhängig von der Größe der betrachteten Software
- Die Fehlerdichte ist unabhängig vom Beobachtungszeitraum
- Die Fehlerdichte unterscheidet sich zwischen Branchen signifikant

Methodik

Die Forschungsarbeit umfasst eine Kombination aus Literaturrecherche und empirischer Analyse von Daten aus verschiedenen Softwareprojekten. Während der Literaturrecherche werden bestehende Studien zu Fehlerdichten betrachtet und der Stand der Forschung herausgearbeitet. Daraus werden die Forschungsfrage und die Hypothesen abgeleitet. Im nächsten Schritt findet die Erhebung von Fehlerdichten aus veröffentlichten Studien unter Berücksichtigung von Faktoren wie Größe der Software, Entwicklungszeitpunkt und Beobachtungszeitraum und Branche statt. Diese Daten werden anschließend analysiert und statistisch untersucht, um Trends und Korrelationen ausfindig zu machen. Schließlich können die Hypothesen bestätigt oder widerlegt, ein Ergebnis abgeleitet und Limitationen der Arbeit betrachtet werden.

Erste Ergebnisse

Eine erste Analyse der Literatur zeigt, dass die Fehlerdichte kontextabhängig ist und einige Limitationen hat. Die Faktoren Entwicklungszeitpunkt, Beobachtungszeitraum, Größe der Software und Branche haben einen Einfluss auf die Fehlerdichte. Die bisher berücksichtigten Daten (siehe Abbildung 1) deuten darauf hin, dass die Fehlerdichte abnimmt, je jünger der Entwicklungszeitpunkt der Software ist. Dies kann durch Verbesserungen des Entwicklungsprozessen und der Güte von Tests erklärt werden. Des Weiteren ist ein Trend erkennbar, dass die Fehlerdichte mit zunehmender Größe der Software abnimmt (siehe Abbildung 2).
Die Fehlerdichte hängt stark vom Beobachtungszeitraum und der Intensität der Nutzung bzw. der Güte des Testens der Software ab. Je länger der Beobachtungszeitraum und je intensiver die Nutzung der Software, desto mehr Fehler werden gefunden. Somit sollte immer auch der Beobachtungszeitraum mit angegeben und berücksichtigt werden, um die Fehlerdichte bewerten und vergleichen zu können. Dieser wird in vielen der untersuchten Studien nicht angegeben.

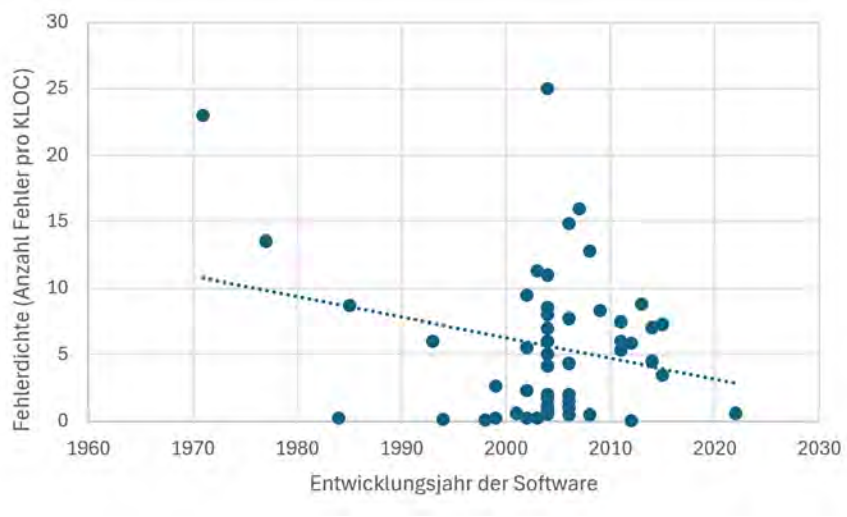


Abb. 1: Streudiagramm mit Fehlerdichte und Entwicklungsjahr der Software [2]

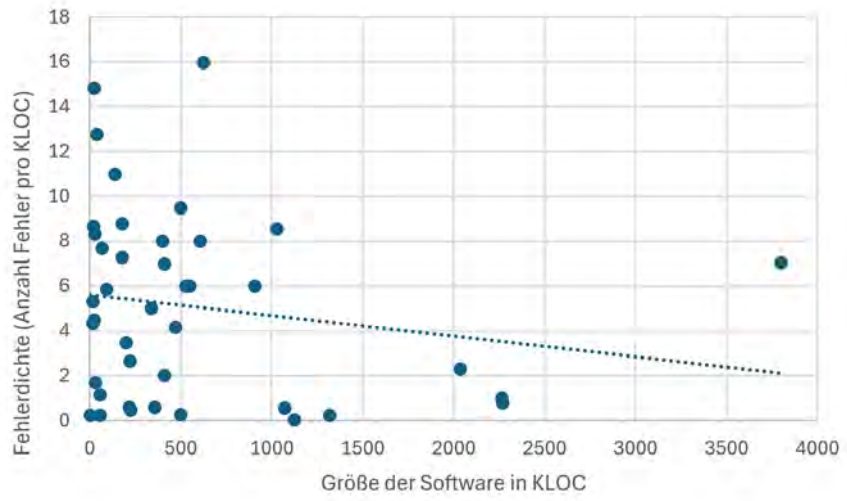


Abb. 2: Streudiagramm mit Fehlerdichte und Größe der Software in KLOC [2]

Ausblick

Die Ergebnisse dieser Arbeit sollen dazu beitragen, ein tieferes Verständnis für die Stärken und Schwächen der Fehlerdichte als Metrik zu schaffen. Im nächsten Schritt werden weitere Daten erhoben und statistisch untersucht. Anschließend sollen Ergebnisse abgeleitet und die Hypothesen beantwortet werden.

Literatur und Abbildungen

[1] Mamdouh Alenezi and Ibrahim Abunadi. Quality of Open Source Systems from Product Metrics Perspective. *International Journal of Computer Science Issues*, 2015.

[2] Eigene Darstellung.

[3] Syed Muhammad Ali Shah et al. An Overview of Software Defect Density: A Scoping Study. In *IEEE*. 19th Asia-Pacific Software Engineering Conference, 2012.

Einsatzanalyse von Künstlicher Intelligenz bezogen auf Geschäftsprozesse in den Bereichen IT-Projektmanagement und Business Development

Sergej Grinko **Thomas Rodach**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Luxoft GmbH, Stuttgart

Einleitung

Künstliche Intelligenz (KI) ist in der heutigen Zeit kein futuristisches Konzept mehr. Es ist eine reale, greifbare Kraft, die das Potenzial hat, grundlegende Tätigkeiten zu verändern. Dabei geht es zudem auch nicht nur um die Automatisierung von Routineaufgaben, sondern eher um die Verbesserung von Service- und Produktqualität sowie die Schaffung neuer Geschäftsmodelle. In der Gesellschaft ist die KI zudem angekommen und hat an Ansehen erlangt. Beginnend mit dem Entsperren des Smartphones durch die biometrische Gesichtserkennung bis hin zu präzisen Empfehlungen durch das Untersuchen des Nutzerverhaltens begleitet uns die KI durch den Alltag. [4]

Aktueller Stand zur deutschen Wirtschaft

Seit 2018 befragt der Digitalverband jedes Jahr 600 Unternehmen mit mehr als 20 Mitarbeitern nach der Planung und Nutzung von KI-Lösungen. Die Umfrage hat ergeben, dass 60 Prozent der Unternehmen die KI als Zukunftstechnologie sehen. 2019 lag der Einsatz von KI in Unternehmen laut der Umfrage bei 2 Prozent. [2] Im Jahr 2024 ist dieser Anteil auf 20 Prozent gestiegen. Dies bedeutet, dass jedes fünfte Unternehmen KI einsetzt. [7]

Relevanz

Der Vorteil an der KI ist, dass es große Mengen an Daten in kurzer Zeit analysieren und bewerten kann, um komplexe Fragestellungen beantworten zu können. Die KI hat zu Beginn des 21. Jahrhunderts die Digitalisierung maßgeblich mitbestimmt. Keine Entwicklung der modernen Gesellschaft wird einen solchen Nachklang hinterlassen, wie die KI. [2]

Problemstellung

Die KI hat das Potenzial, Geschäftsprozesse in fast allen Bereichen grundlegend zu verändern. Durch die Fähigkeit, große Datenmengen zu analysieren, Muster zu erkennen und prädiktive Analysen durchzuführen, ermöglicht es die KI, potenziell effizientere und effektivere Geschäftsprozesse zu entwickeln. Die Bachelorarbeit untersucht die Wirkung von KI auf zwei zentrale Bereiche: IT- Projektmanagement und Business Development.

Aktuelle Herausforderungen für Unternehmen

Von den 354 Unternehmen, die befragt wurden, gaben die Unternehmen, die sich nicht mit KI- Anwendungsfeldern befassen, folgende Gründe an: 44 Prozent der Befragten warten ab, bis sich die Frage klärt, wo sich der Einsatz der KI am meisten lohnt. 46 Prozent gaben an, dass die benötigten finanziellen Ressourcen nicht zur Verfügung stehen, um ein solches Vorhaben in die Wege zu leiten. 40 Prozent der Befragten gaben an, dass sie nicht ausreichende Daten haben. [2]

Teilbereiche der KI

Die Abbildung 1 zeigt die Teilbereiche einer KI. Grundlegend besteht die KI aus dem „Maschinellen Lernen“, bei dem das eigentliche Erschließen der Zusammenhänge aus den Daten geschieht und aus dem zweiten Teilbereich „Deep Learning“. Deep Learning, das wiederum ein Teilbereich des Maschinellen Lernens ist, ermöglicht es der KI, das eigentliche Erlernen von Erkenntnissen. Deep Learning genießt aktuell die größte Aufmerksamkeit in der KI- Forschung. In diesem Zusammenhang werden neuronale Netze verwendet, die sich an der Funktionsweise des menschlichen Gehirns orientieren. [1]

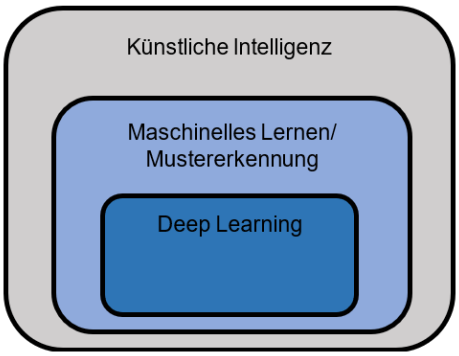


Abb. 1: Teilbereiche der KI [3]

Aus einzelnen künstlichen Neuronen lassen sich wie in Abbildung 2 ersichtlich ganze Netze (kNN) bilden. Die künstlichen Neuronen erhalten ein Inputsignal, welches durch verschiedene Schichten verarbeitet wird, um einen Output zu erzeugen. [3] Um praxisrelevante Funktionen nachbilden zu können, benötigt das kNN mehrere Schichten. Diese ist in drei Arten unterteilt. Die Eingabeschicht nimmt die Werte auf und gibt diese einzeln weiter, die anschließend erneut verarbeitet werden. Darauf folgt die versteckte Schicht. In dieser werden die Werte erneut verarbeitet und schlussendlich an die Ausgabeschicht weitergegeben um einen Output zu erzeugen. [5]

Blackbox Mysterium

Die Funktionsweise der KI ist verflochten und komplex, um Problemstellungen lösen zu können. Die Übersicht zu behalten, ist daher selbst für erfahrene Programmierer eine Herausforderung, geschweige denn zu erläutern, wie die klare Funktionsweise der KI im Detail funktioniert. Die KI-Software besteht aus Millionen von Codezeilen, wobei es die Übersicht zu behalten zudem erschwert. Was demnach zwischen dem Start- und dem Endvorgang geschieht, ist unmöglich nachzuverfolgen. Die KI-Technik ist so komplex, dass die genauen Schritte des Algorithmus nicht nachvollzogen werden können. [6]

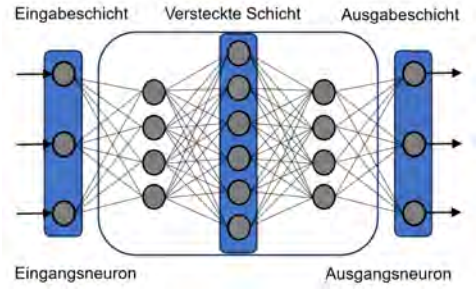


Abb. 2: Tiefes neuronales Netz [5]

Ausblick

Der Ausblick dieser Arbeit hebt die zunehmende Bedeutung der Künstlichen Intelligenz in den Unternehmenskontext hervor. Die Weiterentwicklung und der Einsatz von KI- Modellen werden im Laufe der Zeit eine Schlüsselrolle spielen. Unternehmen müssen sich verstärkt auf die KI-Technologie konzentrieren, um sich Wettbewerbsvorteile sichern zu können.

Literatur und Abbildungen

[1] Tim Cloe. *Erfolgsfaktor Künstliche Intelligenz KI in der Unternehmenspraxis: Potenziale erkennen – Entscheidungen treffen*. Carl Hanser Verlag München, 2020.

[2] Andreas Klug and Jörg Besier. *Trendradar KI Relevante Anwendungsfelder für Unternehmen*. Haufe- Lexware GmbH & Co. KG, 1 edition, 2022.

[3] Patrick Krauss. *Künstliche Intelligenz und Hirnforschung Neuronale Netzte, Deep Learning und die Zukunft der Kognition*. Springer-Verlag GmbH, 2023.

[4] Tawia Odoi. *KI Exzellenz Erfolgsfaktoren im Management Jenseits des Hypes*. Haufe- Lexware GmbH & Co. KG, 1 edition, 2024.

[5] Nils Röttger, Gerhard Runze, and Verena Dietrich. *Basiswissen KI-Testen Qualität von und mit KI-basierten Systemen, Aus- und Weiterbildung zum Certified Tester AI Testing – Foundation Level Specialist nach ISTQB-Standard*. dpunkt. Verlag GmbH, 2024.

[6] Walter Simon. *KI Exzellenz Erfolgsfaktoren im Management Jenseits des Hypes*. BoD – Books on Demand Norderstedt, 2019.

[7] Andreas Streim and Janis Hecker. Erstmals beschäftigt sich mehr als die Hälfte der Unternehmen mit KI. https://www.bitkom.org/Presse/Presseinformation/Erstmals-beschaeftigt-Haelfte-Unternehmen-KI#_, 2024.

Performance Optimierung einer embedded Steuerung durch Vorverarbeiten von OPC UA PubSub Ethernet-Paketen in einem FPGA

Sebastian Haberkern **Walter Lindermeir**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Schneider Electric GmbH, Marktheidenfeld

Einleitung

Open Platform Communication Unified Architecture (OPC UA) ist ein Protokoll im Umfeld der industriellen Kommunikation. Es dient als entscheidendes Protokoll für Industrie 4.0 und ermöglicht eine Kommunikation von der Feldebene bis hin zur Cloud, basierend auf Ethernet TCP/IP. [1]
Zum Austausch der Daten stehen zwei grundsätzliche Architekturen zur Verfügung: Client Server, bei welcher jeder Client eine eigene Verbindung zum Server aufbauen muss, und Publish Subscribe (PubSub), bei der eine Message Oriented Middleware (MOM) den Transport der Nachrichten übernimmt. [3]

Motivation

Bei einer PubSub basierten Kommunikation sendet der Publisher die Nachrichten an die MOM, wobei er nicht weiß, wie viele Subscriber es für diese Daten gibt. Der Subscriber hingegen kennt den Publisher nicht, sondern teilt lediglich der MOM mit, welche Daten er gerne erhalten möchte.

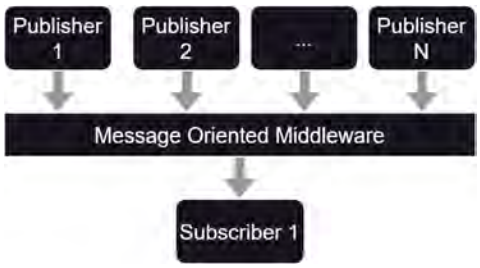


Abb. 1: Schematische Darstellung der PubSub-Kommunikation über die MOM [2]

In einer Beispiel-Anwendung mit N Antrieben und einer Steuerung könnte das wie in Abbildung 1 dargestellt aussehen. Die Kommunikation könnte über IPv4 Multicast-Pakete erfolgen, wodurch es sich um eine Broker-Less MOM handeln würde, die aus der Netzwerk-Infrastruktur wie bspw. Switche besteht, und lediglich die Nachrichten transportiert. Es wird somit für jeden Publisher eine einzelne Nachricht an den Subscriber gesendet.
Da den Daten der Publisher eine eindeutige ID zugeordnet ist müssen diese nicht zwingend in einzelnen Nachrichten an den Subscriber übertragen werden. Dies wäre auch gemeinsam in einer Nachricht möglich. Hierdurch könnte eine Entlastung des Subscribers erfolgen, da dieser für die Daten aller Publisher innerhalb eines Zyklus lediglich eine Nachricht der MOM verarbeiten muss.

Zielsetzung

Das Ziel dieser Arbeit ist das Aggregieren von OPC UA PubSub Nachrichten in einem FPGA. Die Kommunikation erfolgt über UADP IPv4 Multicast-Pakete, welche ein FPGA im Pfad zwischen MOM und Subscriber zwischenspeichern und gesammelt in einer neu erstellten Nachricht an den Subscriber weiterleiten soll. Der Ablauf ist nachfolgend in Abbildung 2 dargestellt.

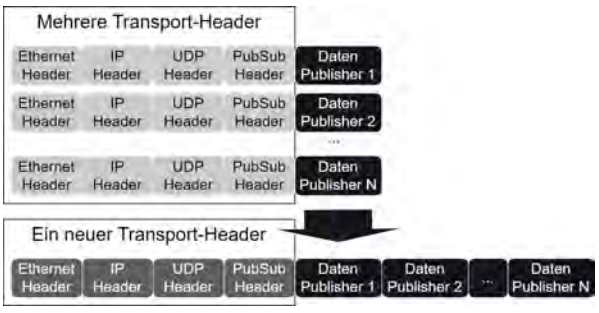


Abb. 2: Schematische Darstellung der Aggregation [2]

Es soll lediglich der in Abbildung 1 dargestellte Pfad berücksichtigt werden – alle anderen Netzwerk-Pakete müssen unverändert weitergeleitet werden.

Testumgebung

Als Grundlage liegt eine Test-Umgebung vor, in der mehrere Publisher zyklisch Daten an einen Subscriber senden. Auf dem Kommunikations-Pfad ist ein FPGA eingebunden, in dessen Blockdesign die Ethernet-Frames an einem AXI-Stream Interface bereit liegen.
Da es sich bei diesem Aufbau um ein komplexes System mit mehreren externen Komponenten handelt, wurde im ersten Schritt eine zweite, virtuelle Test-Umgebung erstellt.
Hierzu wurden zunächst die relevanten Ethernet-Pakete aufgezeichnet. Ein in Verilog neu erstellter IP-Block simuliert anhand dieser aufgezeichneten Ethernet-Pakete eine mögliche Kommunikation an dem AXI-Stream Interface innerhalb des FPGAs. Diese Komponente wurde AXI-S Player genannt. Das erstellte Gegenstück hierzu, das einen AXI-Stream aufzeichnen kann, heißt AXI-S Recorder.
Zur Analyse dieser aufgezeichneten Ethernet-Frames kommt die Software IC-Monitor zum Einsatz, die das Dekodieren und Analysieren diverser industrieller Kommunikations-Protokolle ermöglicht. [4]

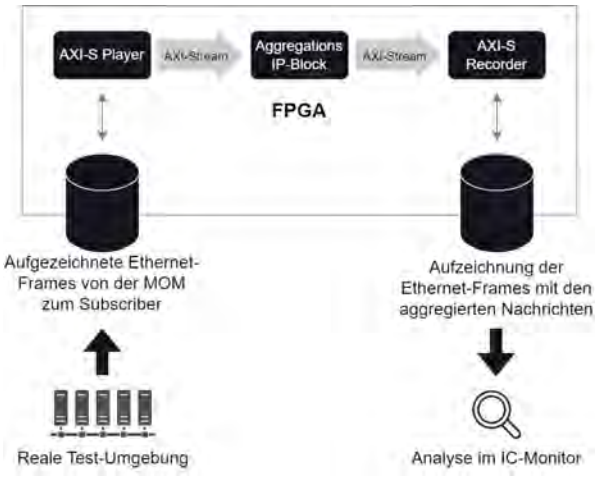


Abb. 3: HIL-Test-Ablauf [2]

Mit diesem Aufbau, der in Abbildung 3 dargestellt ist, lässt sich ein HIL-Test der Ziel-Komponente dieser Arbeit, dem Aggregations IP-Block, durchführen.

Realisierung

Die Entwicklung des Aggregations IP-Blocks erfolgt in Vitis HLS. HLS steht hierbei für High-Level Synthesis und es handelt sich um ein Tool, mit dem es möglich ist aus abstraktem C/C++ Code den Register-Transfer Level (RTL) -Code für die Implementierung im FPGA zu erzeugen.
Die Entwicklung mit diesem Tool erfolgt in der Regel in den folgenden Schritten:

- C/C++ Code: Der Algorithmus wird in einer hohen Abstraktions-Ebene geschrieben. Parallel dazu wird eine C-Testbench erstellt, die die Funktion des Algorithmus verifiziert. [5]
- Erzeugen des RTL-Codes: Aus dem C-Code wird der RTL-Code erzeugt. In einer Simulation des RTL-Codes ist es möglich alle Signale während der Ausführung der C-Testbench taktgenau zu analysieren. [5]
- Optimieren: Mit Hilfe von C directives lässt sich die RTL-Code Generierung beeinflussen. Dadurch kann der Trade-Off zwischen Performance und der benötigten Ressourcen im FPGA beeinflusst werden. [5]

Ausblick

OPC UA bietet außerdem die Möglichkeit, die Daten in einer PubSub Nachricht zu signieren oder zu verschlüsseln. Dies erfolgt immer für alle Daten einer Nachricht gemeinsam. Soll der in Abbildung 2 dargestellte Ablauf bspw. mit verschlüsselten Daten erfolgen, so müssen die einzelnen Daten zunächst im FPGA entschlüsselt und anschließend gemeinsam neu verschlüsselt werden, da jeder Publisher seinen eigenen Schlüssel verwendet.

Literatur und Abbildungen

[1] Wolfgang Babel. *Systemintegration in Industrie 4.0 und IoT*. Springer Vieweg, 2024.

[2] Eigene Darstellung.

[3] OPC Foundation. UA Part 14/ PubSub - Annex B (informative)Client Server vs Publish Subscribe. https://reference.opcfoundation.org/Core/Part14/v104/docs/B, 2024.

[4] Steinbeis Embedded Systems Technologies GmbH. Industrial Communication Protocol Analysis - IC-Monitor. https://www.ic-monitor.com/, 2024.

[5] Advanced Micro Devices Inc. ug1399-vitis-hls-en-us-2024.1. https://docs.amd.com/r/en-US/ug1399-vitis-hls/Design-Principles, 2024.

Einsatz von Künstlicher Intelligenz zur Optimierung von Social Media Marketingstrategien für Start-Ups

Angelina Heine **Catharina Kriegbaum-Kling**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Expatio, Sachsenheim

Einleitung

Laut dem Startup Genome Bericht scheitern etwa 90 % der Start-ups innerhalb der ersten drei Jahre. In einer stark digitalisierten Welt stehen Start-ups vor der Herausforderung, trotz begrenzter Ressourcen wie Zeit, Geld und Arbeitskräfte wettbewerbsfähig zu bleiben und langfristig zu überleben. Marketing spielt hierbei eine entscheidende Rolle, um ein junges Unternehmen bekannt zu machen und auf dem Markt zu etablieren. Insbesondere Social-Media-Marketing bietet eine kostengünstige Möglichkeit, eine breite Zielgruppe zu erreichen, den Bekanntheitsgrad zu steigern und eine aktive Community aufzubauen. Darüber hinaus ermöglicht es direkte Dialoge mit der Zielgruppe, den Aufbau von Vertrauen, die Gewinnung von Neukunden sowie die Einholung von Feedback. [4] Eine starke Unterstützung für diese Prozesse bietet der Einsatz von Künstlicher Intelligenz (KI). KI ist bereits seit Jahren ein fester Bestandteil unseres Alltags, gewann jedoch insbesondere seit der Veröffentlichung von ChatGPT im Jahr 2022 enorme Aufmerksamkeit. Im kreativen Bereich und der Content-Erstellung können KI-Tools Start-ups dabei helfen, Prozesse zu automatisieren, Ressourcen zu sparen und effizienter zu arbeiten – ein entscheidender Vorteil für junge Gründerteams mit limitierten Mitteln. [2]

Theoretischer Hintergrund Social Media Marketing nutzt Plattformen wie soziale Netzwerke, Weblogs und Wikis, um durch gezielte Inhalte die Interaktion mit der Zielgruppe zu fördern und die Markenbindung zu stärken. Im Mittelpunkt stehen relevante und ansprechende Inhalte, die über verschiedene Kanäle verbreitet werden. [4] Künstliche Intelligenz (KI) optimiert und automatisiert diese Prozesse, indem sie menschliches Verhalten simuliert und datenbasierte Entscheidungen trifft. Generative KI, ein spezieller Bereich der KI, erstellt eigenständig neue Inhalte wie Texte, Bilder oder Videos und nutzt erlerntes Wissen, um innovative und kreative Lösungen zu entwickeln (siehe Abbildung 1). Insbesondere für ressourcenarme Start-ups bietet sie ein unverzichtbares Werkzeug für effektives und effizientes Social Media Marketing. Tools wie ChatGPT, Midjourney, Notion AI und Opus Clip ermöglichen die schnelle Erstellung und Anpassung von Inhalten, wodurch Zeit und Kosten gespart werden können. [3]

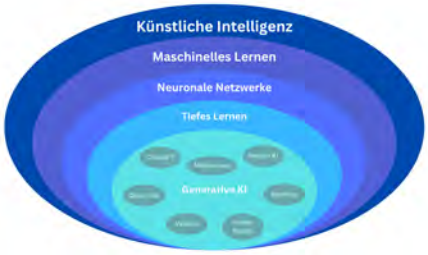


Abb. 1: Einordnung der Generativen KI und ihrer Anwendungen [1]

Zielsetzung Das Ziel dieser Arbeit ist die Entwicklung einer Social-Media-Marketingstrategie für ein Start-up und deren Optimierung durch den gezielten Einsatz von KI. Dies soll insbesondere Start-ups mit begrenzten Ressourcen ermöglichen, ihre Marke effizient und effektiv aufzubauen, bekannt zu machen und in der frühen Phase Kunden zu gewinnen. Eine unstrukturierte Herangehensweise führt oft zu ineffizienter Ressourcennutzung und schlechten Ergebnissen. Daher ist es entscheidend, jede Maßnahme zielgerichtet zu planen und das Vorgehen systematisch zu dokumentieren, um Zeit und Aufwand zu sparen. Die Arbeit umfasst sowohl die theoretische Darstellung der einzelnen Schritte einer Social-Media-Marketingstrategie als auch deren praktische Anwendung in einem Start-up. Dabei werden verschiedene KI-Tools analysiert und die am besten geeigneten in der Praxis eingesetzt. Ziel ist es, praxisnahe und anpassbare Lösungen zu liefern, die Start-ups in der Umsetzung einer erfolgreichen Social-Media-Präsenz unterstützen.

Praxisbeispiel: Expatio

Unternehmensvorstellung Expatio ist ein Ein-

Personen-Start-up, das gegründet wurde, um Migranten in Deutschland bei ihrer Integration zu unterstützen. Es begleitet sie in allen Phasen, von der Planung des Umzugs über die Jobsuche bis hin zur Erfüllung bürokratischer Anforderungen. Da Expatio von einer einzigen Person betrieben wird, stehen der Gründerin nur begrenzte zeitliche und finanzielle Ressourcen zur Verfügung, und es fehlt an Unterstützung. Um diese Herausforderungen zu bewältigen, setzt das Start-up auf eine Social-Media-Marketingstrategie mit KI-Tools. Diese ermöglichen es, den Content-Workflow zu automatisieren, die Community effizient aufzubauen und trotz der geringen Kapazitäten eine starke digitale Präsenz zu etablieren.

Vorgehensweise bei der Entwicklung der SMM Strategie für Expatio Die Entwicklung der Social-Media-Marketingstrategie für Expatio umfasst mehrere Schritte, die aufeinander aufbauen. Zunächst werden SMART-Ziele definiert, um klare Prioritäten wie Reichweitensteigerung, Kundenbindung und Umsatzwachstum zu setzen. Anschließend erfolgt eine Wettbewerbsanalyse mithilfe der SWOT-Methode, um Stärken, Schwächen, Chancen und Risiken der Konkurrenten zu identifizieren. Eine detaillierte Zielgruppenanalyse untersucht demografische, psychografische und verhaltensbezogene Merkmale, die als Basis für die Erstellung von Personas dienen. Darauf aufbauend wird die Plattform-Strategie entwickelt, wobei Plattformen wie YouTube, Instagram und Telegram priorisiert werden, um maximale Reichweite und Engagement zu erzielen. Die Content-Strategie definiert, welche Inhalte erstellt und wie sie präsentiert werden, um die Zielgruppe gezielt anzusprechen, während das Community-Management den aktiven Dialog fördert, Vertrauen aufbaut und Kundenbindung stärkt. Erfolgsmessung und Controlling überwachen die Zielerreichung anhand von KPIs wie Reichweite, Engagement und Umsatz, um datenbasierte Optimierungen vorzunehmen. Abschließend wird die Strategie durch kontinuierliche Anpassungen flexibel auf Markt- und Zielgruppenveränderungen abgestimmt (siehe Abbildung 2).

Abb. 2: Entwicklung einer Social Media Marketingstrategie [1]

Ergebnisse und Diskussion Die Social-Media-Marketingstrategie von „Expatio“, basierend auf der Integration von KI-Tools, lieferte erste wertvolle Einblicke in die Zielgruppe und optimierte die Effizienz der Arbeitsprozesse. Die folgenden Punkte fassen die wichtigsten Ergebnisse und Herausforderungen zusammen:

- Effizienzsteigerung durch KI-Tools: Die Automa-

tisierung von Content-Erstellung, Planung und Analyse führte zu einer erheblichen Zeitersparnis und einem effizienteren Workflow, was besonders in einem Ein-Personen-Start-up von Vorteil ist.

- Anpassung an die Zielgruppe: Inhalte wurden auf die spezifischen Bedürfnisse der Zielgruppe abgestimmt und individuell gestaltet. Dies resultierte in einer verbesserten Zielgruppenansprache und einem höheren Engagement.
- Konsistenter Content-Workflow: Ein zentralisierter Content-Plan sorgte für eine strategische und regelmäßige Veröffentlichung von Inhalten, wodurch die Markenpräsenz gestärkt wurde.
- Erste Ergebnisse der Strategie: Obwohl die Strategie erst kürzlich implementiert wurde und Zeit benötigt, um volle Wirkung zu zeigen, konnte bereits eine verbesserte Interaktionsrate festgestellt werden. Ein deutliches Follower-Wachstum steht jedoch noch aus.

Herausforderungen:

- Begrenzte Ressourcen: Die Herausforderung, alle Aufgaben allein zu bewältigen, erforderte eine präzise Zeit- und Ressourcenplanung.
- Initialer Aufwand: Die Einführung und Anpassung der KI-Tools brachte zu Beginn einen hohen Zeitaufwand mit sich, der jedoch durch langfristige Effizienzgewinne kompensiert wurde.
- Langfristige Zielerreichung: Um die ambitionierten Ziele wie Umsatzsteigerung und Community-Aufbau zu erreichen, ist eine kontinuierliche strategische Optimierung notwendig.

Empfehlungen für weitere Maßnahmen:

- Erweiterung der Plattformnutzung: Aktivitäten auf Plattformen wie LinkedIn und Facebook sollten intensiviert werden, um neue Zielgruppen zu erschließen.
- Community-Building: Durch personalisierte Inhalte, Live-Events und verstärkte Interaktion mit der Community kann die Bindung zur Zielgruppe gestärkt werden.
- Optimierung der Content-Strategie: Der Fokus sollte auf der Anpassung an aktuelle Markttrends und der Erweiterung um spezifische Nischenthemen liegen, um die Zielgruppenansprache weiter zu verbessern.
- Skalierung der Automatisierung: Der Ausbau automatisierter Prozesse kann die Effizienz weiter steigern und die Arbeitsbelastung reduzieren.
- Nachhaltige Content-Strategie: Die Inhalte sollten nicht nur relevant und sinnvoll gestaltet, sondern auch in einer optimalen Länge und mit Mehrwert für die Zielgruppe produziert werden.
- Langfristige Evaluation: Die Strategie sollte regelmäßig überwacht und an neue Anforderungen angepasst werden, um nachhaltige Ergebnisse zu erzielen.

Ausblick

Für Start-ups wie Expatio ist eine effektive Social-Media-Marketingstrategie von zentraler Bedeutung, um trotz begrenzter Ressourcen die Marke bekannt zu machen, eine hohe Reichweite zu erzielen und zielgerichtet mit der Zielgruppe zu kommunizieren. Der Einsatz von KI-Tools unterstützt nicht nur einzelne Prozesse, sondern optimiert die gesamte SMM-Strategie, indem datenbasierte Entscheidungen ermöglicht, Inhalte besser auf die Zielgruppe abgestimmt und Arbeitsabläufe effizienter gestaltet werden. Zukünftig sollten die Strategie weiter verfeinert, nachhaltige Inhalte priorisiert und die Automatisierung ausgebaut werden, um die Markenbekanntheit zu steigern, langfristiges Wachstum zu sichern und die Wettbewerbsfähigkeit zu stärken.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Miriam Löffler and Irene Michl. *Think Content! Strategie. Marketing, Formate*. Rheinwerk, 3 edition, 2024.

[3] Bernhard Wecke. *Wachstum durch den Einsatz Generativer KI*. Springer Gabler, 2024.

[4] Martin Wrobel. *Marketing und Vertrieb für Startups*. Springer Gabler, 2024.

Automatisierte Usability-Test-Generierung durch LLMs: Ein Prototyp für die visuelle Webseiten-Bedienung mit Playwright

Andreas Heinrich **Andreas Rößler**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einführung

Die Benutzerfreundlichkeit und Bedienbarkeit einer Webseite sind zentrale Faktoren für ihren Erfolg. Eine benutzerfreundliche Webseite ermöglicht es Nutzenden, Informationen schnell und einfach zu finden sowie nahtlos mit der Seite zu interagieren. Mit der Integration von Large Language Models (LLMs) eröffnet sich ein innovativer Ansatz zur Analyse der Bedienbarkeit: Diese Modelle sind in der Lage, Webseiten zu interpretieren, Aktionen wie Klicks oder Scrollen auszuführen und somit gezielt die Nutzerführung sowie die Auffindbarkeit von Informationen zu evaluieren und menschliche Interaktion nachzuahmen.
Ansätze wie Steward nutzen OpenAIs LLMs GPT-3.5-Turbo, GPT-3.5-Turbo-16k, GPT-4-Turbo und GPT-4-Vision in Kombination mit dem Browser-Automatisierungstool Playwright, um durch die Analyse von Screenshots und gefiltertem HTML-Code Aufgaben auf Webseiten auszuführen [6].
Durch den systematischen Vergleich verschiedener LLMs und unterschiedlicher Promptansätze in realitätsnahen sowie standardisierten Tests im Rahmen des Mind2Web-Projekts stellt einen bedeutenden wissenschaftlichen Beitrag zur Automatisierung von Usability-Tests dar [2].

Zielsetzung der Arbeit

Ziel der Arbeit ist die Entwicklung eines Prototyps zur systematischen Untersuchung und zum Vergleich von LLMs und unterschiedlicher Promptansätze in realitätsnahen sowie standardisierten Tests im Rahmen des Mind2Web-Projekts zu untersuchen. Dazu werden LLMs wie ChatGPT (OpenAI), Gemini (Google) und Llama (Meta) hinsichtlich ihrer Fähigkeit zur Webseitenanalyse und Nutzerinteraktion verglichen.
Um menschliche Nutzerinteraktionen nachzuahmen, sollen die LLMs, auf Basis von visuellem Input in Form eines Screenshots und einer Liste mit Koordinaten und Größen klickbarer Flächen, Entscheidungen zur Interaktion mit der Webseite treffen, die anschließend mittels Playwright ausgeführt werden. Für die Evaluation und den Vergleich werden standardisierte Tests des Mind2Web-Projekts sowie ausgewählte öffentliche Webseiten herangezogen. Dabei werden neben Erfolg und Misserfolg auch die Anzahl der Requests, die Performance in Bezug auf Zeit, Interaktionen und Kosten der unterschiedlichen Promptansätze und LLMs verglichen.

Technologische Grundlage

Large Language Models

Large Language Models (LLMs) sind ein bedeutender Fortschritt in der künstlichen Intelligenz, die auf tiefen neuronalen Netzwerken basieren und in der Lage sind, eine Vielzahl von Aufgaben zu bewältigen, indem sie unterschiedliche Daten wie Text, Audio, Bilder und Videos verarbeiten [5].
Diese Modelle, wie z.B. ChatGPT oder Gemini, haben das Potenzial, die IT-Branche zu transformieren, indem sie Aufgaben automatisieren und die Effizienz in der Datenverarbeitung steigern [3]. In der Bild- und Textverarbeitung ermöglichen LLMs die Integration multimodaler Daten, wodurch Anwendungen wie visuelles Frage-Antworten realisierbar werden [4].

Playwright

Playwright ist ein modernes Werkzeug für die visuelle Webseiten-Bedienung und Testautomatisierung. Es ermöglicht Entwicklern, Webanwendungen in verschiedenen Browsern zu testen, indem es Benutzerinteraktionen simuliert und Webseiten durchläuft. Playwright bietet eine umfassende Unterstützung für Cross-Browser-Tests, was es zu einer bevorzugten Wahl für Entwickler macht, die sicherstellen möchten, dass ihre Anwendungen in verschiedenen Umgebungen konsistent funktionieren. Playwright bietet eine Entwicklungs-API, die es ermöglicht, Webseiten mit-

tels verschiedenen Programmiersprachen wie Python automatisiert zu bedienen und zu testen, indem sie eine intuitive Schnittstelle für die Interaktion mit Webseiten-Elementen bereitstellt [7].

Architektur des Prototyps

Die Architektur des Prototyps besteht, wie in Abbildung 1 zu sehen, aus vier Microservices, die über RabbitMQ-Message-Queues kommunizieren, sowie einem Reverse-Proxy. Die einzelnen Komponenten sind wie folgt organisiert:

- **Service-Backend-Controller**: Dieser zentrale Service koordiniert die Tests, persistiert alle Daten sowie die zugehörigen Nachrichten der anderen Services und verarbeitet eingehende Ergebnisse. Über Websockets stellt er die Ergebnisse für den Nutzer bereit.
- **Service-Browser-Control**: Verantwortlich für die Steuerung der Webseiten-Interaktionen, einschließlich der Verwaltung und Bereitstellung von Screenshots der Webseiten.
- **Service-LLM-Control**: Zuständig für das Senden von Anfragen an die jeweilige API eines Large Language Models (LLM) und das Empfangen der Antworten.
- **Service-Frontend**: Bereitet die Daten für den Nutzer auf und stellt diese über ein Web-Interface zur Verfügung.

Die Kommunikation erfolgt über Message-Queues, wobei die Antworten von Service-Browser-Control und Service-LLM-Control über eine Antwort-Queue zurück an den Service-Backend-Controller geschickt werden. Dieser verarbeitet die Ergebnisse und überträgt sie in Echtzeit an das Frontend.

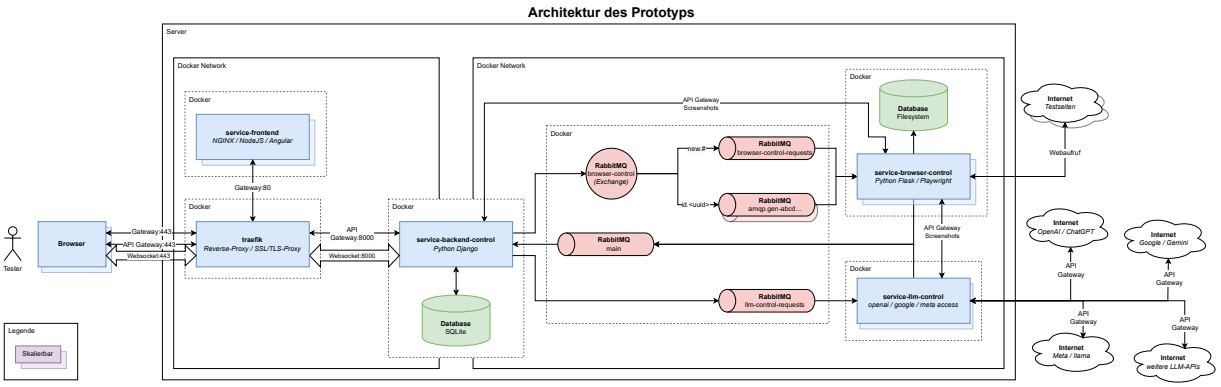


Abb. 1: Architektur des Prototyps [1]

Der Testablauf, wie in Abbildung 2 zu sehen ist, beginnt damit, dass der Tester über das Frontend, das in Abbildung 3 zu sehen ist, eine neue Testsession startet. Dabei wählt er die URL der zu testenden Webseite, das gewünschte LLM sowie das zugehörige Prompt-Set aus. Der Prozess gliedert sich wie folgt:

1. **Initialisierung der Testsession**: Nach Auswahl der Parameter wird über das ausgewählte LLM eine Anfrage formuliert, um den Zweck der angegebenen URL zu ermitteln und zu identifizieren, welche Informationen für den Test sinnvoll wären. Basierend auf dieser Antwort wählt der Tester ein Testziel für die Session aus.
2. **Start des Browser-Services**: Eine Nachricht mit dem Schlüssel new.# wird an den Service-Browser-Control gesendet, der daraufhin einen neuen Browser mit der angegebenen URL startet. Von dieser Webseite wird ein Screenshot erstellt, und es wird eine Liste mit den Koordinaten und den Größen klickbarer Flächen generiert.
3. **Interaktion mit dem LLM**: Der Screenshot sowie die Liste der klickbaren Elemente werden an das LLM mit dem gewählten Prompt-Set weitergeleitet. Das LLM wählt eine geeignete Browser-Interaktion aus oder gibt direkt die gewünschte Information entsprechend dem Testziel zurück.
4. **Ausführung der Interaktion**: Die gewählte Browser-Interaktion wird über eine Nachricht mit dem Schlüssel id.<uuid> an den Service-Browser-Control gesendet. Dies stellt sicher, dass die Nachricht auch bei Skalierung eindeutig dem richtigen Browser-Service zugeordnet wird. Der Browser-Service führt die Interaktion aus und liefert erneut einen Screenshot mit aktualisierten Informationen zu den klickbaren Flächen.

5. **Wiederholung oder Abschluss**: Diese Schritte werden iterativ wiederholt, bis entweder das Testziel erreicht ist oder der Nutzer den weiteren Verlauf nicht mehr freigibt.

Dieser Ablauf gewährleistet eine effiziente und flexible Testautomatisierung, die sowohl Interaktionen als auch Testziele dynamisch steuert.

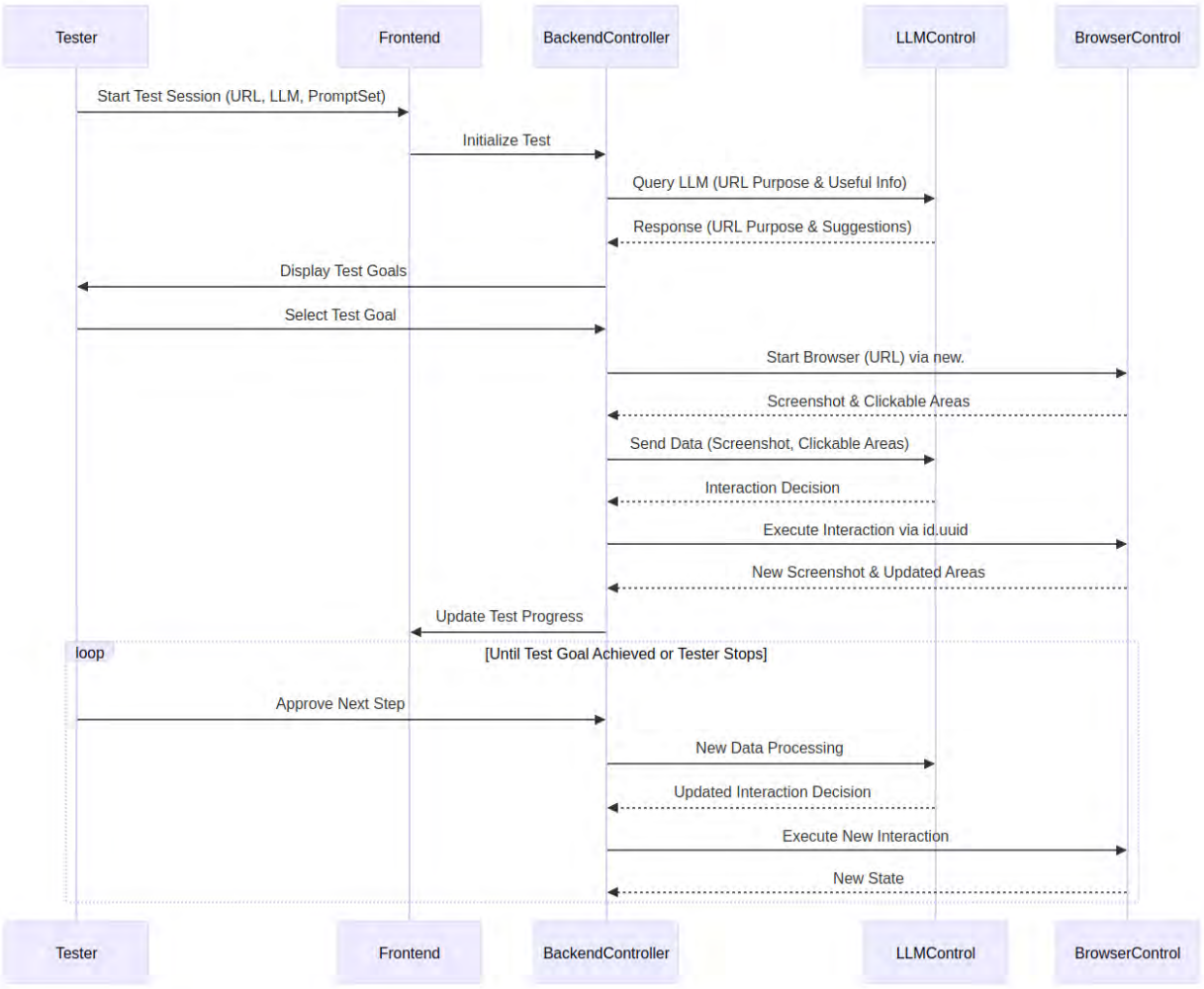


Abb. 2: Sequenzdiagramm eines Testablauf [1]

Erste Erkenntnisse und Ausblick

Das ausgearbeitete Konzept ermöglicht es, wie in Abbildung 3 zu sehen, Daten zu verarbeiten und ChatGPT4o als erstes angebundenes LLM zur Bedienung von Webseiten einzusetzen. Zukünftig sollen weitere APIs für zusätzliche LLMs integriert werden, um die Funktionalität zu erweitern und die Vergleichbarkeit der Ergebnisse zu erhöhen. Zudem werden zusätzliche Browser-Interaktionen implementiert, um dem LLM eine größere Auswahl an Handlungsmöglichkeiten zu bieten. Die Tests basieren auf Teilen des Mind2Web-Projekts, um vergleichbare Ergebnisse zu anderen Publikationen zu erzielen. Abschließend werden die Testergebnisse anhand verschiedener Faktoren wie Performance (Zeit, Interaktionen) und Kosten analysiert und bewertet.

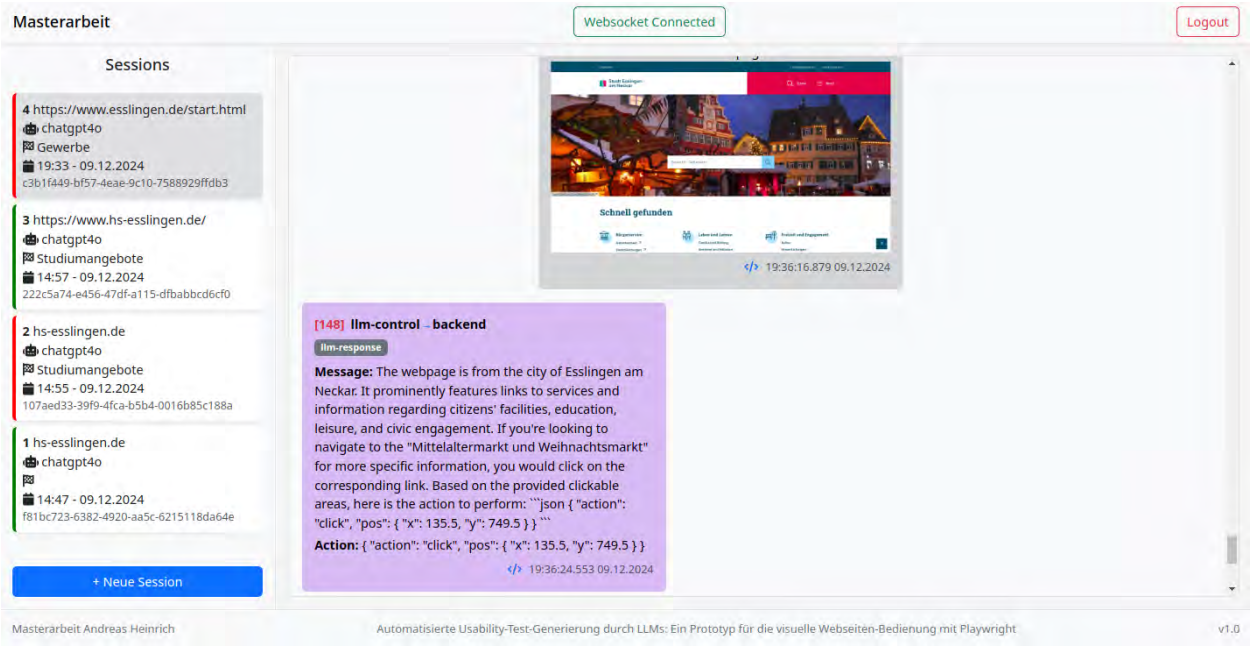


Abb. 3: Angular Frontend zum Verwalten und Visualisiere von Tests [1]

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2Web: Towards a Generalist Agent for the Web. https://osu-nlp-group.github.io/Mind2Web/, 2024.

[3] Haoran Han, Siyao Wu, Jinyao Yang, and Yizhuo Zhao. The foundation, current situation and future prospects of pre-training large language models. *Dean & Francis Academic Publishing*, 2024.

[4] Chris Kelly, Luhui Hu, Bang Yang, Yu Tian, Deshun Yang, Cindy Yang, Zaoshan Huang, Zihao Li, Jiayin Hu, and Yuexian Zou. VisionGPT: Vision-Language Understanding Agent Using Generalized Multimodal Framework. *arxiv.org*, 2024.

[5] Ian Scott and Guido Zuccon. The new paradigm in machine learning – foundation models, large language models and beyond: a primer for physicians. In *Internal Medicine Journal: Volume 54, Issue 5*, volume 54, pages 697–841. Internal Medicine Journal, 5 edition, 2024.

[6] Brian Tang and Kang G. Shin. Steward: Natural Language Web Automation. *arxiv.org*, 2024.

[7] Khevna Vadia, Aryan Thukrul, Priyanka Ghosh Mazumdar, Nikhil Panda, and Ajay Sirsat. Bug Testing Automation with Playwright and a Backend API. In *2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, pages 1867–1870. IEEE, 2024.

Analyse der Eignung eines LIN-Busses zur Ansteuerung eines kinematischen Systems mit zeitkritischen Funktionen

Felix Hintennach **Michael Scharf**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma BOS GmbH & Co. KG, Ostfildern

Motivation

Der Bedarf an kostengünstigen und zuverlässigen Bussystemen steigt in der Automobilindustrie stetig. Besonders für sicherheits- und zeitkritische Anwendungen ist eine zuverlässige Kommunikation zwischen den Komponenten unerlässlich. Der Local Interconnect Network-Bus (LIN-Bus) bietet durch seine einfache Architektur und kosteneffiziente Umsetzung weiterhin eine ideale Lösung für Anwendungen mit geringen Datenraten und moderaten Echtzeitanforderungen. Ziel dieser Arbeit ist es, zu untersuchen, unter welchen Bedingungen zeitkritische Softwarefunktionen, wie der Einklemmschutz, über den LIN-Bus sicher und zuverlässig umgesetzt werden können. Darüber hinaus soll analysiert werden, ob vollintegrierte, softwarelose LIN-Motorsteuerungs-ICs in zukünftigen BOS-Anwendungen eingesetzt werden können.

LIN-Bus

Der LIN-Bus (Local Interconnect Network) ist ein standardisiertes, kosteneffizientes Bussystem, das seit den späten 1990er Jahren vor allem in der Automobilindustrie eingesetzt wird und zur Vernetzung von Steuergeräten dient. Der LIN-Bus eignet sich ideal für Anwendungen mit geringen Datenraten und moderaten Echtzeitanforderungen. Typische Einsatzbereiche umfassen die Steuerung von Klimaanlagen, Fensterhebern und anderen Peripheriesysteme. Technisch basiert der LIN-Bus auf dem Master-Slave-Prinzip, bei dem ein Teilnehmer als Master die Kommunikation steuert und bis zu 15 Slaves integriert werden können. Die Datenübertragung erfolgt dabei sequenziell über Nachrichtenrahmen (Frames) mit einer Geschwindigkeit bis zu 19,2 kbit/s. Dabei zeichnet er sich durch seine hohe Vorhersagbarkeit aus, die durch die Verwendung einer Kommunikationsmatrix (Schedule Table) gewährleistet wird. Diese Matrix definiert die zeitliche Abfolge und den Inhalt der Nachrichten (Frames), die zwischen Master und Slaves ausgetauscht werden. Durch diese Strukturierung ist sichergestellt, dass jedes Frame innerhalb eines festgelegten Zeitrahmens gesendet wird und damit die Anforderungen an die moderate Echtzeitkommunikation erfüllt werden. [4] Zudem ermöglicht das byteorientierte, asynchrone Protokoll des LIN-Bus eine einfache Implementierung auf nahezu jedem Mikrocontroller, da Nachrichten flexibel mit einer Länge von 1 bis 8 Datenbytes übertragen werden können [2]. Diese Eigenschaften machen den LIN-Bus zu einer effizienten Lösung für Anwendungen, die eine zuverlässige, aber einfache Kommunikation erfordern.

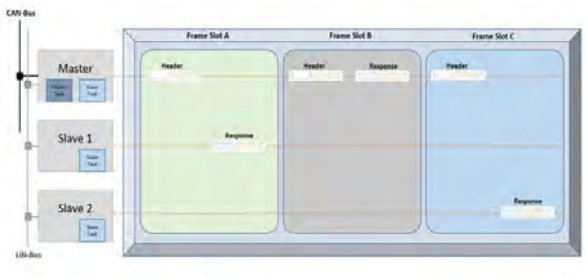


Abb. 1: LIN-Kommunikation [1]

Softwarelose ICs

Ein integrierter Schaltkreis (IC) ist ein vielseitig einsetzbarer Chip, der beispielsweise als Verstärker, Oszillator, Timer, Zähler, Computerspeicher oder Mikroprozessor verwendet werden kann. Es gibt ICs in zwei Ausführung. Die linearen ICs, erzeugen eine stufenlose Ausgangsfunktion und sind somit ideal für Anwendungen als Operationsverstärker. Digitale ICs arbeiten mit binären Logikgattern, die eine festgelegte Anzahl von Zuständen ermöglichen. Diese Eigenschaft macht sie sehr relevant für die Steuerung elektronischer Geräte. [5] Bei BOS GmbH & Co. KG kommen digitale ICs als Schnittstellen zwischen Motortreiber und Bussystem zum Einsatz. Ein typisches Beispiel ist ein elektrisches Fensterrollo, das mit einem DC-Motor betrieben wird. Hier übernimmt der IC die Funktion,

Befehle vom Bussystem zu empfangen und an den Motortreiber weiterzuleiten. Die Logik muss hierbei selbst Entwickelt werden. Softwarelose ICs hingegen integrieren die gesamte Logik des Bussystems bereits hardwareseitig. Für LIN-Bus-Anwendungen werden sie oft als System Basis Chips (SBC) angeboten. Diese Bausteine enthalten bereits sämtliche Funktionen für die Steuerung und die Kommunikation, wodurch der Entwicklungsaufwand erheblich reduziert werden könnte. [3]

Implementierung eines LIN-Busses

Die Implementierung beginnt mit der Entwicklung eines Motorsteuergeräts (LIN-Slave), das die vom LIN-Master empfangenen Befehle ausführt und dessen Rückmeldungen verarbeitet. Dieses Gerät übernimmt nicht nur die hardwarenahe Steuerung des Motors, sondern auch die kontinuierliche Erfassung von relevanten Daten wie Motorposition und Drehrichtung. Darauf aufbauend wird ein LIN-Master-Steuergerät entwickelt. Dieses bildet zentrale Systemfunktionen wie zum Beispiel Initialisierung, Öffnen, Schließen und Einklemmschutz ab. Der LIN-Master steuert die Kommunikation im System, indem er regelmäßig den Status des LIN-Slaves abfragt und darauf basierend Steuerbefehle generiert. Abschließend wird die gesamte Kommunikation zwischen LIN-Master und LIN-Slave validiert. Hierbei gilt es sicherzustellen, dass zeitkritische Funktionen (Beispiel Einklemmschutz) zuverlässig über eine LIN-Schnittstelle abgebildet werden können.

Abb. 2: Systemaufbau [1]

Ausblick

Die Erkenntnisse über den Einsatz eines zuverlässigen Bussystems in Kombination mit softwarelosen ICs können einen bedeutenden Fortschritt für die Entwicklungsprozesse bei BOS darstellen. Durch die Integration dieser Technologie ließen sich nicht nur Entwicklungszeiten reduzieren, sondern auch die Zuverlässigkeit der Anwendungen weiter steigern. Die im Rahmen dieser Untersuchung generierten Messdaten sollen die Entscheidungsfindung unterstützen und eine fundierte Wissensbasis für den zukünftigen Einsatz dieser Technologie schaffen. Dies könnte neue Möglichkeiten in der Automatisierung der Produktion eröffnen.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Andreas Grzemba and Hans-Christian von der Wese. *LIN-Bus: Systeme, Protokolle, Test von LIN-Systemen, Tools, Hardware*. Franzis Verlag GmbH, 2005.

[3] Texas Instruments. TLIN1028-Q1 Automotive LIN 125-mA System Basis Chip (SBC). https://www.ti.com/lit/ds/symlink/tlin1028-q1.pdf?ts=1733490685697, 2022.

[4] Mathias Rausch. *Kommunikationssysteme im Automobil*. Carl Hanser Verlag München, 2022.

[5] Margaret Rouse. Integrierter Schaltkreis (IC, Integrated Circuit). https://www.computerweekly.com/de/definition/Integrierter-Schaltkreis-IC-Integrated-Circuit, 2018.

Prototypisierung einer regionalen Lagenachführung von Bildobjekten

Wei De Huang **Reiner Marchthaler**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Balluff MV GmbH, Oppenweiler

Einleitung

Durch die zunehmende Automatisierung industrieller Prozesse hat sich die industrielle Bildverarbeitung (IBV) in den letzten Jahren zu einem unverzichtbaren Bestandteil der heutigen Industrie entwickelt. In einer Zeit, in der Automatisierung und Digitalisierung Schlüsselrollen in der Industrie einnehmen, spielt die präzise und zuverlässige Bildverarbeitung eine entscheidende Rolle bei der Qualitätssicherung und Effizienzsteigerung von Produktionsprozessen. Hochentwickelte Bildverarbeitungssysteme sind vor allem in Branchen wie der Automobilindustrie, dem Maschinenbau und der Medizintechnik unverzichtbar [3].

Ein zentraler Aspekt dieser Bildverarbeitungssysteme ist die automatische Lokalisierung von Werkstückträgern, welche die Grundlage für die eigentlichen Sichtprüfaufgaben von Objekten bildet. Insbesondere die zuverlässige Anwesenheitskontrolle und Inspektion von Objekten auf Werkstückträgern erfordert eine exakte X/Y-Positionierung und Orientierung des Werkstückträgers in jedem aufgenommenen Bild [2]. Aktuell existiert bei der Balluff MV GmbH eine Anwendungssoftware mit grafischer Benutzeroberfläche (GUI), siehe Abbildung 1, die es dem Benutzer ermöglicht, mit einem Tool (Presence Check) eine Anwesenheitskontrolle anhand eines Referenzbildes zu trainieren.

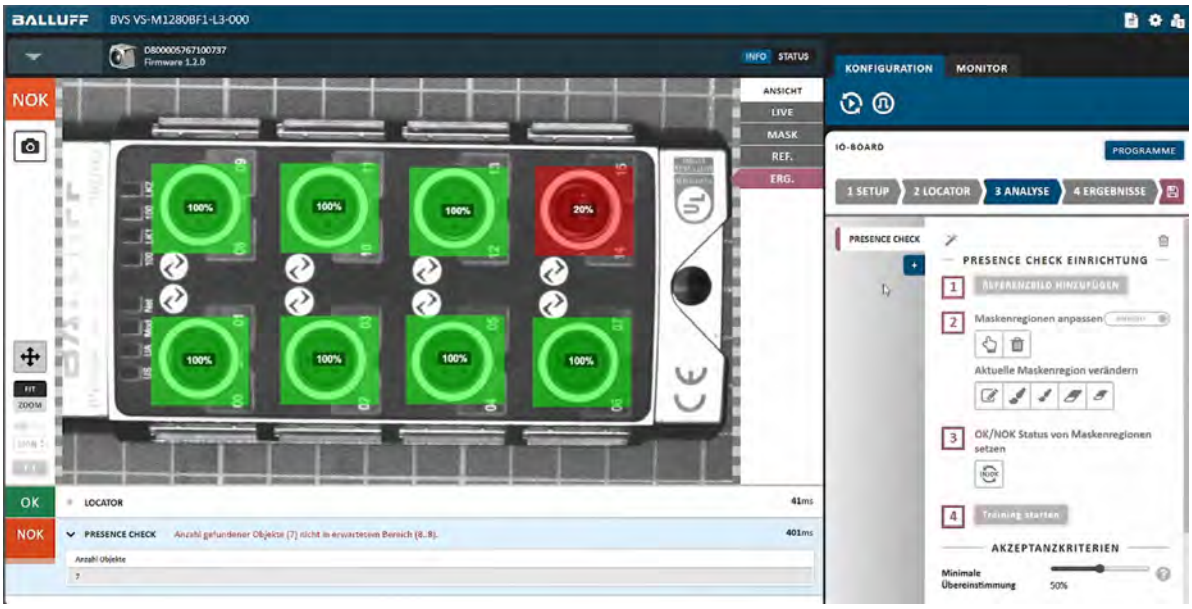


Abb. 1: Benutzeroberfläche des BVS VisionSensors [1]

Der Presence Check prüft die Anwesenheit des trainierten Objekts innerhalb von Regionen, die vom Benutzer definiert wurden, siehe Abbildung 1. Bisherige Methoden und Techniken stoßen jedoch häufig an ihre Grenzen, insbesondere wenn mehrere gleichartige Objekte in komplexen Anordnungen auf Werkstückträgern geprüft werden sollen. Schwierig wird es, wenn die Positionen und/oder Orientierungen dieser

einzelnen Objekte leicht von der definierten Anordnung abweichen. In solchen Fällen kann die vorhandene Anwesenheitskontrolle Schwierigkeiten haben, zuverlässig zu erkennen, ob alle Objekte vorhanden sind. Dies liegt daran, dass die Übereinstimmung mit den trainierten Referenzbildern abnimmt, sobald die Positionen und/oder Orientierungen von den bekannten Referenzpositionen abweichen.

Abb. 2: Beispiel einer Anwesenheitskontrolle von beweglichen Objekten im Werkstückträger [1]

Ziel der Arbeit

Konkret soll für den BVS VisionSensor das kombinierte Verfahren zur Lokalisierung und Anwesenheitskontrolle mehrerer gleichartiger Objekte auf einem Werkstückträger um einen Zwischenschritt erweitert werden. Der Balluff VisionSensor ist eine intelligente Industriekamera mit integriertem Betriebssystem und benutzerfreundlicher GUI, die in der Lage ist, Bildverarbeitungsaufgaben autonom durchzuführen. Durch die Entwicklung erweiterter prototypischer Algorithmen soll die Robustheit und Zuverlässigkeit der bestehenden Anwesenheitskontrolle (Presence Check) verbessert werden. Das Ergebnis soll in der Lage sein, Objekte trotz geringfügiger Änderungen ihrer X/Y-Position und Orientierung im Produktionsbetrieb zuverlässig zu erkennen. Gleichzeitig muss sichergestellt bleiben, dass fehlende oder falsche Objekte nicht als korrekt erkannt werden. Außerdem darf die Ausführungsgeschwindigkeit nicht zu sehr verlangsamt werden. Das Verfahren soll auch nach der Verbesserung in möglichst vielen Anwendungsfällen ohne Benutzereingriff auskommen.

Auswahl eines geeignetes Matchingverfahren

Zunächst wurden die Rahmenbedingungen und Anforderungen an die Lagenachführung identifiziert. Zur Auswahl eines geeigneten Matchingverfahrens wurden verschiedene Verfahren analysiert. Letztendlich wurde das formbasierte Matching als das geeignetste Verfahren ausgewählt, da es robust gegenüber Verdeckung und Clutter, also zusätzlichen oder fehlenden Kanten im Bild, ist und die Transformationsparameter X/Y-Position und Orientierung zurückgeben kann. Das formbasierte Matching ist ein Verfahren zur Objekterkennung in der Bildverarbeitung, das auf der Form der Konturen eines Objekts basiert. Es wird verwendet, um Objekte in Bildern zu finden, indem ein Modell des Objekts erstellt und dieses Modell dann in anderen Bildern gesucht wird [4] .

Umsetzung der Lagenachführung

Der Prozess der Lagenachführung mit dem formbasierten Matching besteht aus fünf Hauptschritten:

1. Auswahl eines Objekts im Referenzbild: Ein Objekt im Referenzbild wird ausgewählt, das als Vorlage für das Modell dient.
2. Erstellung eines geeigneten Modells: Ein Formmodell wird mit dem ausgewählten Objekt erstellt. Dieses Modell beschreibt die Form der Konturen des Objekts.
3. Training des Modells: Das erstellte Modell wird trainiert, indem geeignete Konturen und Kanten des Objekts im Template gesucht und für das Modell verwendet werden.
4. Finden von Objektinstanzen: Mit dem trainierten Modell werden in neuen Bildern Objekte gesucht, die dem Modell entsprechen. Die Position, Orientierung und Skalierung der gefundenen Objekte werden zurückgegeben.
5. Korrektur der Objektposition: Abschließend wird die Position des gefundenen Objekts mittels affiner Transformation korrigiert.

Ergebnis

Ein Prototyp der Lagenachführung wurde in dieser Arbeit mit einem formbasierten Matchingverfahren erfolgreich umgesetzt. Insgesamt stellt der in dieser Arbeit entwickelte Prototyp eine solide Grundlage für die weitere Entwicklung des Presence Checks dar. Die Lagenachführung ist besonders effektiv in Prüffällen, in denen die Prüfobjekte entweder im Referenzbild oder zur Laufzeit in ihrer Lage und Orientierung variieren. Die Integration des Prototypen in die Produktionslinie des BVS VisionSensors ist realistisch. Abbildung 3 zeigt ein Beispielergebnis einer Anwesenheitskontrolle von bewegten Objekten mit der Lagenachführung.

Abb. 3: Beispielergebnis einer Anwesenheitskontrolle mit Lagenachführung [1]

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Christian Demant et al. *Industrielle Bildverarbeitung: wie optische Qualitätskontrolle wirklich funktioniert*. Springer, 3 edition, 2011.

[3] Verband Deutscher Maschinen und Anlagenbau. VDMA Branchenführer Machine Vision. https://www.vdma.org/viewer/-/v2article/render/17002727, 2021.

[4] Carsten Steger et al. *Machine Vision Algorithms and Applications*. Wiley-VCH, Berlin, 2 edition, 2018.

Analyse von Mobilitätsdaten für die Prognose des zukünftigen Mobilitätsbedarfs

Medjen Izairi **Mirko Sonntag**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Anwendungszentrum KEIM des Fraunhofer Instituts für Arbeitswirtschaft und Organisation (IAO), Esslingen

Einleitung

Die zunehmende Urbanisierung stellt Städte weltweit vor enorme Herausforderungen. Umweltprobleme wie Luftverschmutzung und steigende CO -Emissionen erfordern innovative Lösungen, um die Lebensqualität in urbanen Räumen zu sichern und gleichzeitig ökologische Nachhaltigkeit zu gewährleisten. Nach Angaben der Vereinten Nationen werden im Jahr 2030 etwa 60 % der Weltbevölkerung in Städten mit mindestens einer halben Million Einwohnern leben, was den Druck auf die bestehenden Verkehrssysteme und die Infrastruktur erhöht [8]. Mit Inkrafttreten der Verordnung über Elektrokleinstfahrzeuge mit Lenk- und Haltestange am Straßenverkehr wurde im Juni 2019 in Deutschland ein wichtiger Schritt zur Mikromobilität gemacht. Sie umfasst die Nutzung von kleinen betriebenen Elektrofahrzeugen (z. B. Tretroller), die sich besonders für kurze Strecken eignen . Diese Fortbewegungsmittel können den ÖPNV ergänzen, indem sie als umweltfreundliche Alternative individuellen Bedürfnissen dienen, insbesondere für die sogenannte „letzte Meile“. Mikromobilität kann dazu beitragen, Verkehrsstaus zu reduzieren, Emissionen zu senken und verschiedene Verkehrsmittel besser zu vernetzen [4] [1]. Um jedoch die Mobilitätsbedürfnisse der Zukunft präzise zu erfüllen, ist eine datengestützte Mobilitätsanalyse unverzichtbar. Durch die Auswertung von Mobilitätsdaten können Verkehrsströme besser verstanden, Engpässe identifiziert und Maßnahmen für eine optimierte Fahrzeugverteilung abgeleitet werden. Diese Analysen bilden die Grundlage für die Prognose zukünftiger Mobilitätsbedarf und die Entwicklung von Strategien, die den Mobilitätsbedürfnissen gerecht werden [5] [2].

Problemstellung

Im Rahmen des Projekts „CHANGE“ des Anwendungszentrum KEIM wird eine dezentrale Ladeinfrastruktur für Fahrzeuge der Mikromobilität bereitgestellt, um die ökologische und ökonomische Effizienz der Mobilitätsanbieter zu fördern. Ein zentraler Bestandteil ist das Anreizsystem, das Nutzer über die App der Mobilitätsanbieter dazu motivieren soll, ihre Fahrzeuge an bestimmten Standorten abzustellen, um eine optimale Fahrzeugverteilung und eine Kundenzufriedenheit zu gewährleisten. Durch das gezielte Abstellen der Fahrzeuge an strategischen Punkten können CO -Emissionen reduziert werden, da weniger Servicefahrten nötig sind [6]. Für ein effektives Anreizsystem ist es relevant zu berücksichtigen, welche Faktoren die Nutzung von Mikromobilitätsfahrzeugen beeinflussen. Beispielsweise können Buchungen von Faktoren wie Stoßzeiten, Tageszeiten, Wetterbedingungen oder der Nähe zu bestimmten Points of Interest (POIs). Gleichzeitig existieren geografische Bereiche, in denen Nutzer die App zwar öffnen, jedoch keine Buchungen tätigen. Daher ist es entscheidend, für das Anreizsystem vorherzusagen, wo und wann in der Zukunft mit hoher Wahrscheinlichkeit eine Buchung stattfinden wird.

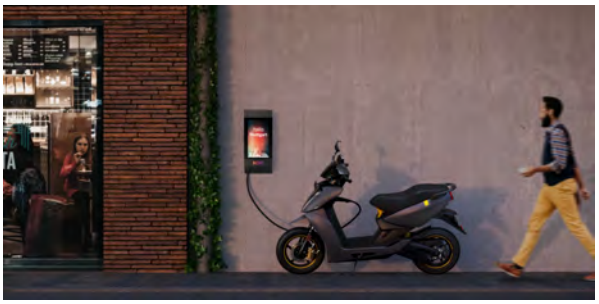


Abb. 1: E-Scooter an einer Ladestation [6]

Zielsetzung

Das Ziel der Arbeit ist die Entwicklung eines Modells zur Vorhersage zukünftiger Buchungen (zum Zeitpunkt t+1), um eine gezielte und effektive Incentivierung von Nutzern zu ermöglichen. Hierzu werden relevante

Parameter definiert, zeitliche Abschnitte analysiert und die Daten entsprechend aufbereitet, um ein geeignetes Modell auszuwählen und zu trainieren, das präzise Vorhersagen ermöglicht.

Vorgehen

Das Vorgehen orientiert sich am CRISP-DM-Modell (Cross Industry Standard Process for Data Mining), das einen strukturierten Ablauf in sechs Phasen bietet. Die Analyse basiert auf anonymisierten Mobilitätsdaten eines Mikromobilitätsanbieters. Diese umfassen Buchungsdaten mit Informationen wie Start- und Endpunkt einer Buchung sowie geografische Daten zum Öffnen und Schließen der App (ab 2020). Ergänzend dazu werden Points of Interest (POI) und Wetterdaten berücksichtigt. Die sechs Phasen des Modells sind:

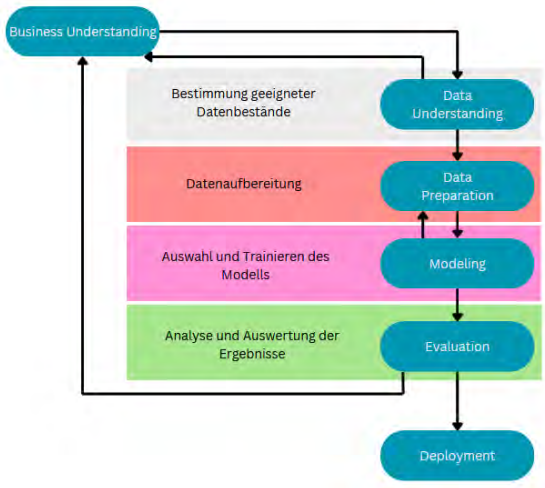


Abb. 2: Das CRISP-DM Prozessdiagramm [3]

1. **Business Understanding**In dieser Phase sollen grundsätzlich die Ziele und Anforderungen festgelegt werden. Dabei wird die Projektplanung erstellt, in der sowohl die zeitlichen Vorgaben als auch die verfügbaren Ressourcen berücksichtigt werden.
2. **Data Understanding**: Hier erfolgt die Sammlung, Exploration und Überprüfung der benötigten Daten, um ein umfassendes Verständnis ihrer Qualität und Struktur zu gewinnen.
3. **Data Preparation**: Es werden die Daten aufbereitet, bereinigt und in ein konsistentes Format gebracht, um einen abschließenden Datensatz zu erstellen, der für das Training des Modells genutzt wird.
4. **Modeling**: Auf Basis des vorbereiteten Datensatzes wird das geeignete Modell ausgewählt und trainiert, um Buchungsprognosen zu erstellen.
5. **Evaluation**: Das Modell wird anhand geeigneter Metriken wie Genauigkeit oder Fehlerrate überprüft. Dies stellt sicher, ob das Modell die Projektziele erfüllt und verlässliche Ergebnisse liefern kann.
6. **Deployment**: Nachdem die Qualität des Modells überprüft wurde, können die gewonnenen Erkenntnisse genutzt werden. In dieser Phase wird das entwickelte Modell in die Praxis übertragen. [7]

Ausblick

Bei einer Integration des Modells in die Geschäftsprozesse könnten die Ergebnisse genutzt werden, um weitere Anwendungen wie die Verbesserung der Nutzererfahrung oder die Minimierung von Leerfahrten zu entwickeln. Langfristig könnte dies dazu beitragen, die Prognosegenauigkeit zu erhöhen, die Systemeffizienz zu steigern und Mikromobilitätsangebote nachhaltiger und kundenorientierter zu gestaltet.

Literatur und Abbildungen

[1] D. Banister. The sustainable mobility paradigm. *Transport Policy*, pages 73–80, 2008.

[2] C. Chen and C.-H. Chao. Urban Mobility Data Analysis Using Machine Learning for Prediction and Optimization. *Journal of Big Data Research*, pages 1–12, 2020.

[3] Eigene Darstellung.

[4] Deutsches Institut für Urbanistik Difu. Mikromobilität im Kontext städtischer Mobilitätskonzepte. *Deutsches Institut für Urbanistik*, 2, 2021.

[5] I. Docherty, G. Marsden, and J. Anable. The governance of smart mobility. *Transportation Research Part A: Policy and Practice*, 2018.

[6] IAO Frauenhofer. CHANGE. https://www.keim.iao.fraunhofer.de/de/projekte/change.html, 2024.

[7] M. Hurst, M. Wentzien, and D. Schmalzried. *Erklärbare künstliche Intelligenz im CRISP-DM-Prozess*. Springer, 2023.

[8] Department of Economic United Nations and Social Affairs. The World's Cities in 2016: Data Booklet. *Department of Economic and Social Affairs, Population Division*, 2016.

Konzeption eines standardisierten Infrastruktur- und Deployment-Frameworks für verteile Systeme basierend auf der Analyse bestehender Infrastruktur

Alessa Jakobs **Jörg Nitzsche**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma adesso SE, Dortmund

Einführung

In der heutigen IT-Welt ist die Bereitstellung von Softwareanwendungen in verschiedenen Produktionsumgebungen eine allgegenwärtige Herausforderung. Diese Umgebungen unterscheiden sich oft in ihren Konfigurationen, den verwendeten Tools und den Deployment-Workflows. Die vorliegende Arbeit befasst sich mit einer Cloud-Infrastruktur, in der Tools wie GitHub Actions und Jenkins zum Einsatz kommen. Trotz eines hohen Automatisierungsgrades führt diese Heterogenität zu fragmentierten Prozessen, die die Wartung, Skalierung und vor allem die Zuverlässigkeit der Deployments erschweren. Ziel dieser Arbeit ist die Entwicklung eines standardisierten Infrastruktur- und Deployment-Frameworks auf Basis von Kubernetes. Dieses Framework soll als Blueprint dienen, um die Deployments zu vereinheitlichen, die Versionierung von Microservices zu vereinfachen und Fehlerbehebungen effizienter zu gestalten. Ein Schwerpunkt liegt auf der Implementierung eines GitOps-Ansatzes, um automatisierte, nachvollziehbare und sichere Deployment-Prozesse zu ermöglichen.

Problemstellung und Herausforderungen

Die Analyse des bestehenden Systems hat gezeigt, dass innerhalb desselben unterschiedliche Konfigurationen und Workflows zum Einsatz kommen. Diese Inkonsistenz führt zu Problemen bei der Standardisierung und Automatisierung der Deployment-Prozesse. Die Verwendung verschiedener Tools führt zu fragmentierten Workflows, was die Wartung der Systeme erschwert. Ein weiteres Problem ist die Verwaltung verschiedener Versionen von Microservices, die sich durch die Heterogenität der Umgebungen als schwierig erweist. Hinzu kommen langwierige Genehmigungsprozesse, die Deployments verzögern, da verschiedene Teams und Stakeholder involviert sind. Die unterschiedlichen Versionen der Microservices auf Entwicklungs-, Produktions- und Staging-Umgebungen führen zu Problemen beim Testen neuer Features und der Behebung von Fehlern. Gleichzeitig ist es aber notwendig, gezielt unterschiedliche Versionen in verschiedenen Umgebungen zu testen, um die Kompatibilität sicherzustellen. Im Fehlerfall sind Rollbacks auf eine frühere Version oft schwierig und zeitaufwendig, da die manuelle Rückgängigmachung von Änderungen in verschiedenen Umgebungen komplex ist. Schließlich können die Konfigurationen der verschiedenen Umgebungen im Laufe der Zeit voneinander abweichen ("Konfigurationsdrift"). Dies erschwert die Reproduzierbarkeit von Problemen und die Sicherstellung einer konsistenten Umgebung für die Anwendung.

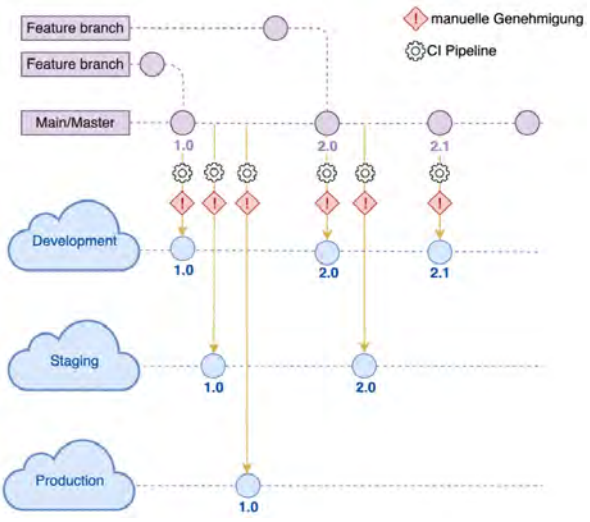


Abb. 1: Beispiel Deployment-Ablauf eines Microservices [2]

Analyse der bestehenden Infrastruktur

Im Rahmen der Arbeit wurde die bestehende Kubernetes-basierte Infrastruktur umfassend analysiert. Der Schwerpunkt lag dabei auf der Struktur des Kubernetes-Clusters, dem Aufbau und Einsatz der CI/CD-Pipelines sowie den Strategien für Rollout- und Rollback-Management. Zunächst wurde die Cluster-Konfiguration untersucht, einschließlich der Anzahl und Konfiguration der Nodes, der Verwendung von Namespaces zur logischen Trennung von Ressourcen und der Implementierung von Ingress-Regeln für den externen Zugriff auf die Services im Cluster. Des Weiteren wurden die CI/CD-Pipelines evaluiert, die auf GitHub Actions und Jenkins basieren. Hierbei wurde analysiert, wie die Workflows aufgebaut sind und welche Automatisierungsgrade erreicht werden. Ein weiterer wichtiger Punkt war die Analyse der Netzwerksicherheit, wobei der Ingress-Controller, die TLS-Terminierung und die bestehenden Netzwerk-Policies im Detail betrachtet wurden. Die Ergebnisse der Analyse zeigen, dass die Prozesse zwar bereits automatisiert sind, jedoch durch die Vielfalt an Tools, Workflows und Konfigurationen schwer zu standardisieren sind. Ein einheitliches Framework auf Basis von Kubernetes und GitOps bietet hier eine geeignete Lösung.

GitOps: Ein Lösungsansatz

GitOps ist eine Methodologie und Praxis, bei der Git-Repositories als „Single Source of Truth“ für die Bereitstellung von Infrastruktur und Anwendungen dienen. Es basiert auf den Grundsätzen der DevOps-Kultur und bietet eine strukturierte Herangehensweise, um den gesamten Entwicklungs- und Betriebsprozess zu optimieren. GitOps basiert dabei auf drei wesentlichen Prinzipien: Git dient als alleinige Quelle der Wahrheit, alle Elemente werden als Code behandelt und sämtliche Operationen werden über Git-Workflows ausgeführt. Ein wesentliches Merkmal von GitOps ist der deklarative Ansatz. Dabei wird der gewünschte Zustand eines Systems durch deklarative Beschreibungen festgelegt, die dann von einer GitOps-Engine überwacht und umgesetzt werden. Die GitOps Engine ist für den CD-Teil der CI/CD-Pipeline zuständig. [4]
Eine Methode GitOps zu implementieren ist pull-basiertes GitOps. Hierbei ziehen Software-Tools (wie ArgoCD oder Flux) automatisch die in Git gespeicherten Zustandsdefinitionen und wenden diese auf das System an. Dieses Modell stellt sicher, dass Konfigurationsänderungen bei Bedarf nicht direkt am System vorgenommen werden, sondern stets über definierte Genehmigungs- und Änderungsprozesse erfolgen. Dadurch wird Drift – also die Abweichung zwischen dem tatsächlichen und dem gewünschten Zustand – erkannt und korrigiert, bevor Probleme auftreten. [4] [1]

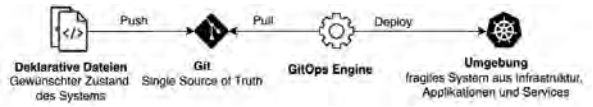


Abb. 2: Vereinfachte Pull-basierte GitOps-Pipeline [2]

Vorteile des GitOps-Ansatzes

Verbesserte Nachvollziehbarkeit: Änderungen an der Anwendungsumgebung werden durch Anpassungen der Konfigurationen im Git-Repository vorgenommen. Dies führt zu einer detaillierten Historie aller Änderungen am gewünschten Zustand.
Vereinfachte Rollbacks: Durch das Zurücksetzen einzelner Commits im Git-Repository kann ein System problemlos in jeden vorherigen Zustand zurückversetzt werden.
Erhöhte Sicherheit: GitOps vereinfacht die Verwaltung, da Operationen innerhalb der Umgebung durchgeführt werden können und nur der Zugriff auf das Git-Repository und die Image-Registry erforderlich ist. Dadurch benötigen Entwickler keinen direkten Zugriff auf die Umgebung.
Toolunabhängigkeit: Das GitOps-Design ist nicht an bestimmte Tools gebunden. Es kann problemlos an verschiedene Toolsets angepasst werden und bietet die Flexibilität, Tools nach Bedarf auszuwählen und zu kombinieren.
Vergleich verschiedener Umgebungen: Die deklarative Speicherung von Konfigurationen in GitHub-Repositorys vereinfacht die Nachverfolgung von Unterschieden zwischen Umgebungen während der Entwicklung, beim Testen und in der Produktion.
Integrierte Backups: Die Verwendung von Git-Repositorys zur Speicherung des Umgebungszustands dient als integriertes Backup für den Cluster-Zustand. Dies gewährleistet die Erhaltung der Datenintegrität bei Störungen in Kubernetes. [3]

Ausblick und Proof of Concept

Um die Praktikabilität des entwickelten Frameworks zu demonstrieren, wird ein Proof of Concept (PoC) durchgeführt. Dieser PoC dient dazu, den GitOps-Ansatz in einer kontrollierten Umgebung zu testen und erste Erkenntnisse zur Umsetzung zu gewinnen. Ziel des PoCs ist es, ein skalierbares, wartbares und konsistentes Deployment-Framework zu entwickeln, das die Deployment-Prozesse vereinfacht und die Zuverlässigkeit erhöht. Langfristig soll der GitOps-Ansatz die Betriebseffizienz durch einheitliche Prozesse erhöhen, die Nachvollziehbarkeit durch eine zentrale Änderungshistorie verbessern und Fehler und Ausfälle durch automatisierte Rollbacks und CI/CD-Pipelines minimieren.

Literatur und Abbildungen

[1] Florian Beetz and Simon Harrer. GitOps: The Evolution of DevOps? In *Engineers, Institute of Electrical and Electronics Software*, volume 4, pages 70–75. Engineers, Institute of Electrical and Electronics, 2022.

[2] Eigene Darstellung.

[3] Saikiran Reddy J, Durga Prashanth Padal S, Albert Mayan J, Catharine A, and J. Jeslin Shanthamalar. Efficient Application Deployment: GitOps for Faster and Secure CI/CD Cycles. In *2024 International Conference on Advances in Modern Age Technologies for Health and Engineering Science (AMATHE)*, pages 1–7. Engineers, Institute of Electrical and Electronics, 2024.

[4] Natale Vinto and Alex Soto Bueno. *GitOps Cookbook*. O'Reilly, 2023.

Design and Evaluation of a Declarative State Management for AUTOSAR Adaptive in the Context of In-Vehicle Container-Orchestration

Tolgahan Kandemir **Mirko Sonntag**

Department of Computer Science and Engineering, Esslingen University

Work carried out at Vector Informatik GmbH, Stuttgart

Abstract

The automotive industry's shift towards Software-Defined Vehicles (SDVs) requires innovative approaches to managing system states in complex, containerized environments. This work proposes a new declarative state management system for the AUTOSAR Adaptive Platform, inspired by concepts and design patterns of state-of-the-art state managers such as Kubernetes. The central research question is: How can Kubernetes-inspired state management patterns be integrated into AUTOSAR to enable dynamic deployments and simplified state management? The proposed solution leverages Kubernetes' declarative model to optimize the deployment of new applications while ensuring system stability and scalability. The results demonstrate that the concept allows for a robust, dynamic deployment process and provides an efficient framework for managing states in an SDV context. This work is significant as it addresses the growing need for flexible, scalable solutions in the evolving automotive industry, offering valuable insights into the future development of state management for SDVs.

Introduction

The relevance of state management within the AUTOSAR Adaptive Platform is becoming increasingly significant as the automotive industry progresses toward the development of SDVs. As vehicles evolve into more flexible, software-driven systems, the need for robust and dynamic state management solutions becomes critical. This shift requires new approaches to managing the lifecycle of applications and system states in a way that supports the flexibility and adaptability expected of SDVs - without redefining state management during design phase for each deployment [6].

This study seeks to address the following research question: How can state management capable of supporting SDV requirements be defined within the context of AUTOSAR Adaptive, and what considerations must be made regarding a declarative approach and enabling dynamic deployments? By exploring these questions, the work aims at identifying key requirements for state management solutions that not only align with the needs of SDVs but also facilitate the seamless integration of new applications and services in dynamic, containerized environments. A declarative state management approach is considered central to achieving these goals, offering a way to simplify and standardize the deployment process while ensuring system stability and scalability.

Theoretical Foundations

AUTOSAR Adaptive: AUTOSAR Adaptive is a modern software platform tailored to SDVs, supporting dynamic updates, service-oriented architectures, and real-time operations. It enables vehicle manufacturers to deploy flexible and scalable solutions to meet the demands of connected and autonomous systems [4].

Kubernetes and State Management: Kubernetes is an open-source platform for orchestrating containerized applications, known for its scalability, resilience, and declarative approach to resource management. Kubernetes ensures that systems autonomously achieve and maintain a user-defined desired state, simplifying complex system operations [5].

State Management in Vehicles: State management in vehicles governs transitions between operational states, ensuring coordinated behavior across systems. It is essential for safety and efficiency, especially in modern vehicles requiring dynamic updates and real-time responsiveness [3].

Function Groups in AUTOSAR Adaptive: Function Groups are logical containers for adaptive applications in AUTOSAR, managing deployment, states, and lifecycles [3]. They can be compared

to Kubernetes pods, serving as units for managing resources and encapsulating functionalities.
Declarative vs. Imperative State Management: Declarative state management defines the desired end state, allowing the system to determine the transition path, as seen in Kubernetes. In contrast, imperative approaches specify exact steps, offering control but at the cost of scalability and simplicity [8].

Proposed Concept

During the conceptualization phase, three approaches for a platform-based state manager were developed:

1. **Standardize State Handling and States**: States are standardized to enable unified and simplified state management for both machine and application lifecycles. This approach proposes multiple state handling options for applications, selectable via configuration during deployment.

2. **Custom State Handling with OEM-Specific Controller**: In this approach, the OEM provides a unique controller for each deployment, which includes the state handling definition. This method offers greater flexibility but is more complex, allowing the OEM maximum freedom in defining state transitions and states itself.

3. **Hybrid Approach**: This approach combines the previous two. A set of standard state handlings is predefined to cover default use cases, but the OEM is given the option to implement custom state handling for use cases that exceed the standard behavior. An optional controller, managed by the state management platform module, can be supplied externally for this purpose.

Of the three approaches considered, the third approach was selected for implementation as it effectively combines the benefits of standardized state management for deployments with the flexibility to accommodate custom state management for specific use cases. This hybrid solution ensures both simplicity and adaptability, addressing a broader range of requirements and use cases. The developed approach and the corresponding concept for a platform-based State Manager are illustrated in figure 1.

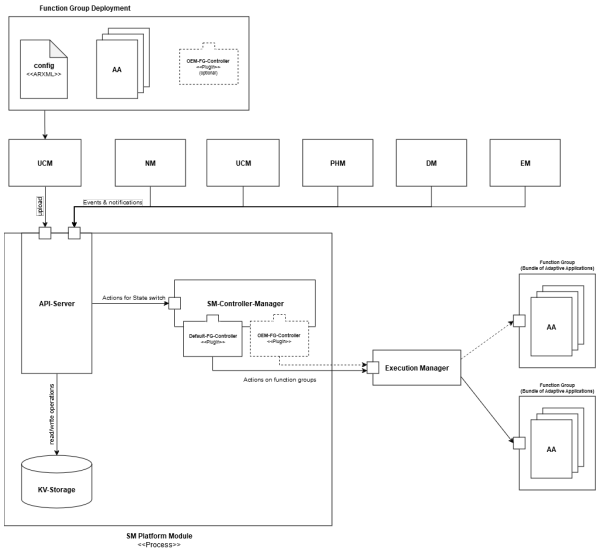


Fig. 1: Concept of a Platform State Manager [7]

The State Manager is restructured into a platform module to decouple application-specific logic from platform-specific functionality, enhancing flexibility and scalability. As seen in the architecture, three key entities from Kubernetes - *API-Server*, *KV-Storage*, and *Controller-Manager* - along with the principle of a controller, are adopted and applied in this concept. The API-Server processes events and determines necessary state changes using mappings stored in the KV-Storage. The Controller-Manager, designed as a plugin-based system, allows OEMs to dynamically deploy custom controllers with state machine logic, enabling a hybrid approach to state management.
In the reconciliation loop, the API-Server identifies required state transitions triggered by events, delegates actions to the Controller-Manager, and forwards them to controllers, which execute the state machine logic. These changes are then applied to function groups via the execution manager, ensuring consistent transitions and efficient resource allocation. This approach integrates Kubernetes concepts with standardized state handling and OEM-specific customization, making it well-suited for modern software-defined vehicles.

Challenges in AUTOSAR R23.11 and State Management in Kubernetes

In Kubernetes, state management is approached by describing states and injecting them into the system, enabling autonomous state transitions via the control plane without the need to explicitly define these transitions. In contrast, the AUTOSAR Adaptive Platform relies on events sent from the Adaptive Platform Modules to the State Manager, which then triggers state transitions. While external input signals in Kubernetes are directly treated as states, in

AUTOSAR Adaptive, they are events, leading to state changes in both environments.
A further challenge arises in the definition of deployment units. In Kubernetes, containerized applications are deployed together within pods, whereas in AUTOSAR Adaptive, multiple applications are grouped into function groups [2]. To simplify the modeling process, this study assumes that a pod in Kubernetes corresponds to a function group in AUTOSAR Adaptive, capable of hosting multiple adaptive applications.
Another challenge lies in the controller logic and execution manager. In Kubernetes each controller is responsible for managing a specific resource type and therefore is directly linked to it [1]. However, within the AUTOSAR Adaptive Platform, controllers, such as the function group controller, are indirectly linked to the resources. The execution manager is responsible for starting and monitoring all applications. One potential approach would be to incorporate controller logic directly into the execution manager, assigning each function group its own execution manager. However, due to the implementation constraints of the execution manager, this is not feasible. Instead, the controller logic, including state transitions, must be integrated into the platform state manager, resulting in all actions being routed through a single execution manager. While integrating the controller logic into the execution manager would lead to a cleaner design, the constraints of the current execution manager implementation prevent this approach.

Conclusion and Future Outlook

This work introduces a restructured state manager concept for AUTOSAR Adaptive, addressing the previous limitation where application and platform logic were tightly coupled, making dynamic deployments nearly impossible. By implementing a centralized platform state manager, applications can now be deployed dynamically without requiring a redesign of the underlying platform. The concept adopts a Kubernetes-inspired approach, centralizing state management and delegating state machine logic to individual controllers, each associated with a specific function group.
While a fully declarative and autonomous state management system, as seen in Kubernetes, cannot be entirely achieved in AUTOSAR Adaptive due to its inherent reliance on external components, this work demonstrates how declarative principles can be applied. By standardizing states and transitions through predefined state machines, and using external events to trigger autonomous transitions, the proposed approach achieves a significant level of autonomy. These results are a critical step towards enabling a Software-Defined Vehicle (SDV) architecture, delivering a fully functional centralized state management system that was previously absent in AUTOSAR releases, including latest R24.11.
Future improvements could focus on restructuring the execution manager to better align with Kubernetes principles, embedding controller logic directly with the resource (e.g., the function group). Currently, AUTOSAR constraints require centralized state management via the execution manager, but future developments could integrate this logic into individual execution managers for each function group, enhancing modularity and scalability while moving closer to a Kubernetes-like model.

References and figures

[1] Kubernetes Authors. Controllers. https://kubernetes.io/docs/concepts/architecture/controller/, 09 2024.

[2] Kubernetes Authors. Pod Lifecycle. https://kubernetes.io/docs/concepts/workloads/pods/pod-lifecycle/, 04 2024.

[3] Adaptive AUTOSAR. Specification of State Management. https://www.autosar.org/fileadmin/standards/R23-11/AP/AUTOSAR_AP_SWS_StateManagement.pdf, 11 2023.

[4] Adaptive AUTOSAR. Adaptive Platform. https://www.autosar.org/standards/adaptive-platform/, 11 2024.

[5] Tim Bannister. Kubernetes Documentation. https://kubernetes.io/docs/home/, 04 2024.

[6] Harald Proff and Elmar Pritsch. Software Defined Vehicle. https://www.deloitte.com/de/de/Industries/automotive/analysis/software-defined-vehicles.html, 10 2024.

[7] Own representation.

[8] Benoit Ruiz. Declarative vs Imperative. https://dev.to/ruizb/declarative-vs-imperative-4a7l, 10 2021.

Der Einsatz von Künstlicher Intelligenz in datenfokussierten Anwendungen mit Java: Frameworks, APIs und ihre Auswirkungen auf die Softwareentwicklung

Joey Kiss **Andreas Rößler**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Novatec Consulting GmbH, Leinfelden-Echterdingen

Einleitung

Die Integration von Künstlicher Intelligenz (KI) hat in den letzten Jahren nicht nur die Art und Weise verändert, wie Anwendungen entwickelt werden, sondern auch neue Herausforderungen für Entwickler*innen geschaffen. Java, eine der etabliertesten Programmiersprachen, bietet durch die Verbindung verschiedene Frameworks und APIs, zu sehen in 1, die Möglichkeit moderne KI-Modelle, also Large Language Models (LLMs), effizient einzubinden. Dieser Artikel untersucht, wie sich diese Technologien in datenfokussierten Anwendungen nutzen lassen und welche Auswirkungen sie auf die Softwareentwicklung haben.

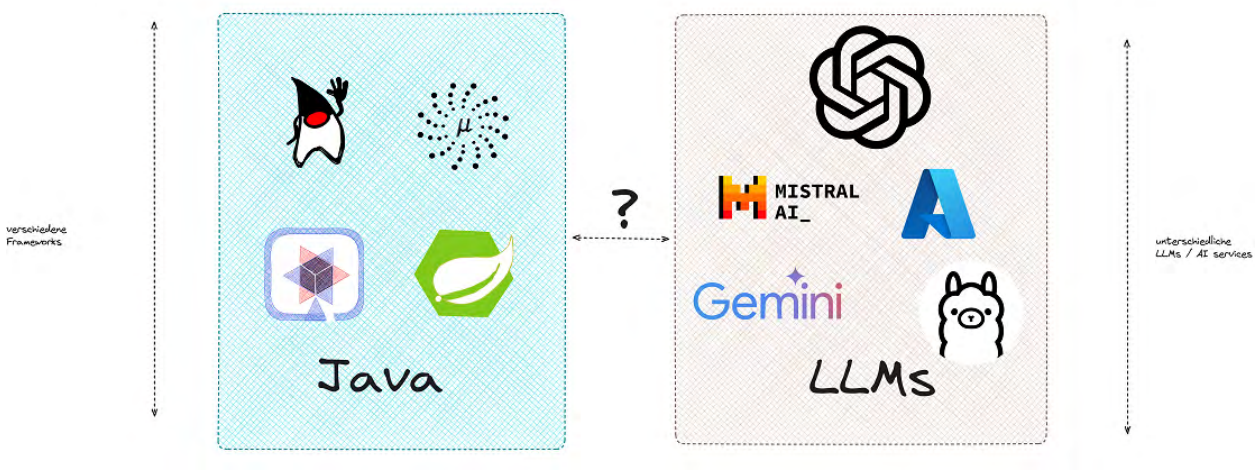
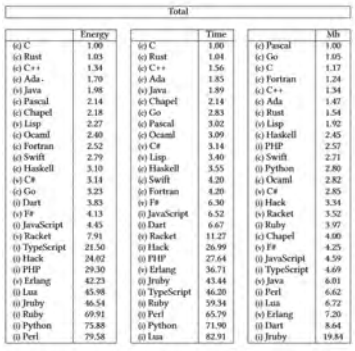


Abb. 1: Verbindung von Frameworks und LLM APIs [1]

Motivation

Die zunehmende Verfügbarkeit von LLMs hat das Potenzial, Softwareentwicklungsprozesse zu vereinfachen und neue Möglichkeiten für datenfokussierte Anwendungen zu eröffnen. Dabei zeigt sich Java als effiziente Plattform: Vergleichende Studien belegen, dass Java in Bezug auf Energieverbrauch und Ausführungszeit sehr performanter ist als Python. Konkret benötigt Java nur 1,98 Energieeinheiten im Vergleich zu 75,58 Einheiten bei Python und reduziert die Ausführungszeit von 71,90 auf 1,89 Zeiteinheiten in 2. Dies macht Java zu einer attraktiven Alternative.



Total

	Energy		Time		Mb
(c) C	1.00	(c) C	1.00	(c) Pascal	1.00
(c) Rust	1.03	(c) Rust	1.04	(c) Go	1.05
(c) C++	1.34	(c) C++	1.56	(c) C	1.17
(c) Ada	1.70	(c) Ada	1.85	(c) Fortran	1.24
(v) Java	1.98	(v) Java	1.89	(c) C++	1.34
(c) Pascal	2.14	(c) Chapel	2.14	(c) Ada	1.47
(c) Chapel	2.18	(c) Go	2.83	(c) Rust	1.54
(v) Lisp	2.27	(c) Pascal	3.02	(v) Lisp	1.92
(c) Ocaml	2.40	(c) Ocaml	3.09	(c) Haskell	2.45
(c) Fortran	2.52	(v) C#	3.14	(i) PHP	2.57
(c) Swift	2.79	(v) Lisp	3.40	(c) Swift	2.71
(c) Haskell	3.10	(c) Haskell	3.55	(i) Python	2.80
(v) C#	3.14	(c) Swift	4.20	(c) Ocaml	2.82
(c) Go	3.23	(c) Fortran	4.20	(v) C#	2.85
(i) Dart	3.83	(v) F#	6.30	(i) Hack	3.34
(v) F#	4.13	(i) JavaScript	6.52	(v) Racket	3.52
(i) JavaScript	4.45	(i) Dart	6.67	(i) Ruby	3.97
(v) Racket	7.91	(v) Racket	11.27	(c) Chapel	4.00
(i) TypeScript	21.50	(i) Hack	26.99	(v) F#	4.25
(i) Hack	24.02	(i) PHP	27.64	(i) JavaScript	4.59
(i) PHP	29.30	(v) Erlang	36.71	(i) TypeScript	4.69
(v) Erlang	42.23	(i) Jruby	43.44	(v) Java	6.01
(i) Lua	45.98	(i) TypeScript	46.20	(i) Perl	6.62
(i) Jruby	46.54	(i) Ruby	59.34	(i) Lua	6.72
(i) Ruby	69.91	(i) Perl	65.79	(v) Erlang	7.20
(i) Python	75.88	(i) Python	71.90	(i) Dart	8.64
(i) Perl	79.58	(i) Lua	82.91	(i) Jruby	19.84

Abb. 2: Zeit-, Energie- und Speicherverbrauch verschiedener Programmiersprachen [2]

Zielsetzung

Ziel dieser Untersuchung ist es, die praktischen Möglichkeiten und Herausforderungen der Integration von KI in Java-Anwendungen durch Frameworks wie Spring AI, Quarkus LangChain4j und andere zu analysieren. Neben der Bewertung der Entwicklungsreife und Konfigurationsflexibilität stehen auch spezifische Features im Fokus: Chat Completion, Text-to-Image, Text-to-Speech und Structured Outputs. Zukünftige Features: Tools/Function Calling und Retrieval-Augmented Generation (RAG). Die Fragen die sich dazu stellen werden in 3 visuell gezeigt.

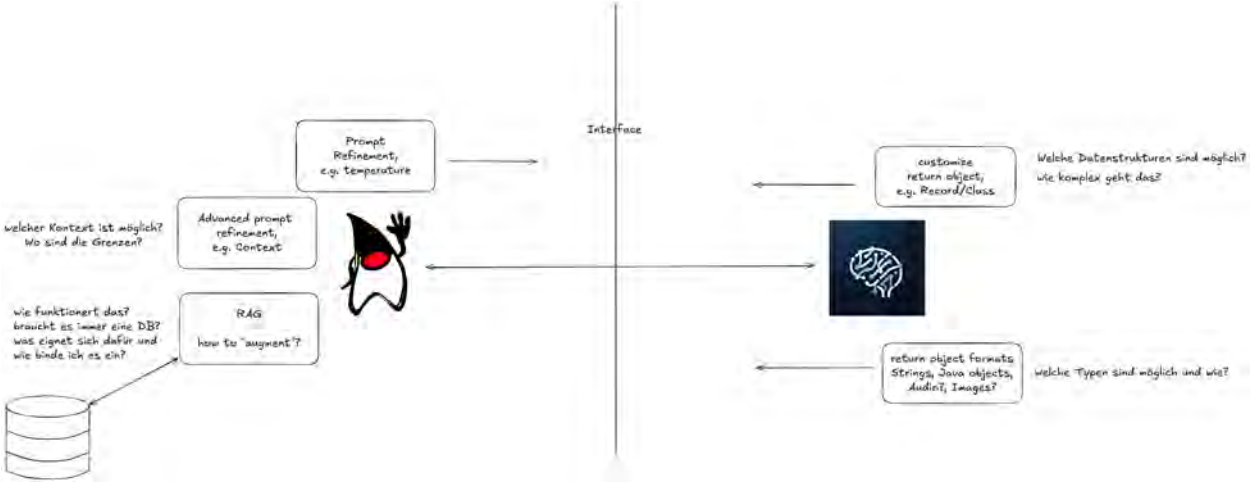


Abb. 3: Interaktion von Java mit AI [1]

Vorgehensweise

Um diese Untersuchung durchzuführen, werden jeweilig Projekte zu jeder LLM dependency erstellt und getestet, das gilt für beide Frameworks. Auswahl der Frameworks Die Analyse umfasst folgende Java-Frameworks und Tools: Spring AI: Ein etabliertes Framework mit umfassender Unterstützung für KI-Modelle und APIs. Quarkus LangChain4j: Ein modernes Framework mit besonderem Fokus auf dynamische Modellintegration. Auswahl der Large Language Models Folgende LLMs und Plattformen ausgewählt: ChatGPT, Azure mit OpenAI Integration, Ollama und Mistral AI Die Untersuchung umfasst eine strategische Auswahl an LLMs mit unterschiedlichen Einsatzszenarien. ChatGPT/OpenAI wurde aufgrund seiner Bekanntheit und führenden Technologien ausgewählt, Azure OpenAI Integration repräsentiert die Unternehmens-Perspektive, Ollama deckt den Bedarf an lokalen Modellen ab, und Mistral AI ergänzt die Auswahl durch seine Zugänglichkeit mit kostenlosen API-Keys.

Ergebnisse

Ergebnisse der bisherigen Arbeit werden in jeweils die Erkenntnisse der Frameworks und LLM spezifischen Dependencies gewertet, Framework-Vergleich Spring AI: Charakteristika: Konsistente Antwortgenerierung, umfangreiche Dokumentation. Herausforderungen: Gelegentliche Timeout-Probleme. Quarkus LangChain4j: Charakteristika: Hohes Level an Abstraktion, einfache Konfiguration, durch Quarkus langchain4j Implementation sehr identischer Code bei Benutzung verschiedener LLMs Herausforderungen: Inkonsistente Antwortgenerierung, komplexeres Parsing durch hohes Level an Abstraktion sowie auch Timeout Probleme LLM-Vergleich Die Untersuchung umfasste vier unterschiedliche Large Language Models mit spezifischen Stärken und Herausforderungen. ChatGPT präsentierte sich als umfassendste Lösung mit den meisten Features und breiter Funktionalität, gleichzeitig aber auch mit komplexen Integrationsanforderungen. Die Azure OpenAI Integration punktete durch robuste Sicherheitsfunktionen und präzise Skalierungsmöglichkeiten, stand jedoch vor Herausforderungen bei EU-Deployments, insbesondere bei bildgenerierenden Modellen, die alternative Standorte wie Schweden oder Australien erfordern. Ollama überzeugte durch seine Offline-Fähigkeiten und lokale Modell-Integration, zeigte jedoch deutliche Performanz-Unterschiede je nach Hardware - von 30 Sekunden Antwortzeit auf Laptops bis zu 2 Sekunden auf leistungsstarken Computern. Mistral AI bot flexible Modellkonfigurationen und gute Entwicklerdokumentation. Feature-Analyse Die Untersuchung der verschiedenen LLM-Implementierungen offenbarte interessante Ergebnisse über die Funktionalität verschiedener Features:
Chat Completion war über alle untersuchten Plattformen konsistent. Bei Quarkus zeigte sich besonders, dass der Implementierungscode für jeden LLM nahezu identisch blieb - lediglich die Dependency variierte

zwischen den verschiedenen Modellen.
Structured Output mit Java Records zeigte präzise Ergebnisse. Die Implementierung verwendete eine List of Strings als Rückgabetyp, wobei die LLMs präzise nur die gewünschte Liste ohne zusätzlichen erläuternden Text zurückgaben. Diese Genauigkeit und Zielgerichtetheit war ein wesentliches Merkmal der Tests und funktionierte bei allen getesteten LLMs und Frameworks konsistent.
Text-to-Image zeigte sich selektiver. Nur ChatGPT und Azure OpenAI ermöglichten eine erfolgreiche Integration, wobei Spring und Quarkus mit Timeout-Problemen kämpften nach 10 Sekunden.
Text-to-Speech war am limitiertesten. Lediglich ChatGPT in Verbindung mit Spring erlaubte eine erfolgreiche Implementierung.
Diese Analyse verdeutlicht die unterschiedlichen Stärken und Herausforderungen bei der LLM-Integration in verschiedene Java-Frameworks.

Diskussion

Die Untersuchung hat gezeigt, dass jedes Framework spezifische Stärken und Schwächen hat. Während Spring AI durch Stabilität punktet, bietet Quarkus LangChain4j höhere Abstraktion. Azure OpenAI überzeugt mit Sicherheitsfeatures, während Ollama und Mistral AI sich durch einfache Konfiguration, fast kostenfreie Nutzung und lokal gehostete Optionen auszeichnen. Die noch ausstehenden Funktionen Tools/Function Calling und RAG werden diese Frameworks weiter bereichern, insbesondere für komplexere Anwendungsfälle.

Ausblick

Die Integration von KI in datenfokussierte Anwendungen mit Java zeigt vielversprechende Ansätze. Zukünftige Entwicklungen könnten sich auf folgende Aspekte konzentrieren: Automatisierung und Vereinfachung der Modellkonfiguration, erweiterte Unterstützung für multimodale Anwendungen und Verbesserte Fehlerbehandlung und Timeout-Vermeidung.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] David Kessel. The fastest computer languages and which programming language is both fast and energy-efficient? A comparison of 27 languages. https://blog.csdn.net/weixin_32540087/article/details/118776660, 2024.

Optimierung einer unternehmensweiten Lernplattform mit besonderem Fokus auf Cloud Computing - Analyse, Konzeptionierung und Handlungsempfehlung

Isabell Kitzberger **Anke Bez**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Dr. Ing. h.c. F. Porsche AG, Stuttgart

Einleitung

Die globale Digitalisierung hat dazu geführt, dass eine Vielzahl von Informationen jederzeit und nahezu überall zugänglich ist. Die scheinbar unbegrenzte Verfügbarkeit von Wissen und Daten sorgt dafür, dass der Druck auf Individuen und Organisationen steigt, sich kontinuierlich weiterzubilden. Die kontinuierliche Weiterbildung wird damit zur essenziellen Voraussetzung, um den Herausforderungen der dynamischen Wissensgesellschaft zu begegnen und die eigene Wettbewerbsfähigkeit langfristig zu sichern. Im bisherigen Corporate-Learning-Bereich erfuhr das sogenannte E-Learning insbesondere durch die Auswirkungen der Corona-Pandemie einen zusätzlichen Verbreitungsboost. Dabei verblieb auch nach der Pandemie die Verbreitung und Nutzung von E-Learning auf einem hohen Niveau. 2024 berichteten 87,9 % befragter Unternehmen im DACH-Raum, E-Learning bereits im Einsatz zu haben. Weitere 9,3 % gaben an, den Einsatz von E-Learning derzeit zu planen. [5]
Der Top-Player im Bereich Education Technology (EdTech) für Corporate Learning sind und bleiben auch weiterhin Lernmanagementsysteme (LMS), bestätigt durch 84,2 % befragter DACH-Unternehmen. Trotz ihres jahrzehntelangen Einsatzes bleibt die Bedeutung von LMS durch aktuelle Neuerungen und Innovationen, wie z.B. Learning Experience Plattformen (LXP), unangefochten. [5] Mit dem zunehmenden Wandel, dass Cloud Computing in den letzten Jahren zu einem zentralen Bestandteil der IT-Infrastruktur zahlreicher Unternehmen wurde, zeichnet sich auch im Bereich der LMS ein klarer Trend von On-Premises-Lösungen hin zu cloudbasierten Ansätzen ab. [1]

Lernmanagementsysteme und Cloud Computing

Im Allgemeinen handelt es sich bei einem LMS um ein System, das der umfassenden administrativen und didaktischen Unterstützung von Lernprozessen dient. [6]
Genauer umfasst ein LMS dabei:

- die Automatisierung von Schulungsprozessen und Schulungsfunktionen,
- Registrierungs- und Verwaltungstools,
- die Verwaltung von Fähigkeiten und Aufzeichnungen,
- den Zugang zu Kursunterlagen
- und Programmierschnittstellen zu Anwendungspaketen. [3]

Ein grundlegender Aufbau für ein LMS kann dabei wie folgt dargestellt werden (siehe Abbildung 1):

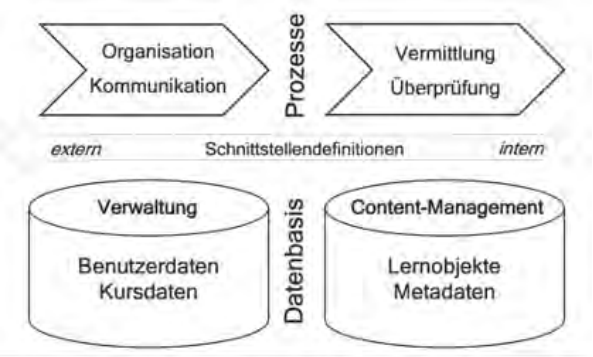


Abb. 1: Vereinfachter LMS-Aufbau [4]

Ein einheitlicher Funktionsumfang eines LMS lässt sich nicht klar definieren, da die angebotenen Systeme in ihrem Funktionsspektrum erheblich variieren können. Umfangreichere LMS können so bedarfsorientiert Funktionen integrieren, die ursprünglich in anderen Kategorien von EdTech-Tools angesiedelt waren, beispielsweise Content-Erstellungsfunktionen,

die typischerweise in Autorentools zu finden sind. Dennoch zeichnen die ursprünglichen Kernfunktionen eines LMS - die Organisation und Administration von Lernprozessen - diese weiterhin aus. Die Vorteile von E-Learning bzw. LMS-Einsatz im Corporate-Learning sind zahlreich und umfassen u.a.:

- die zentrale Organisation und Verwaltung von Lernprozessen und Schulungsmaßnahmen;
- die örtliche und zeitliche Flexibilität bei der Durchführung von Schulungsmaßnahmen;
- die Skalierbarkeit von Schulungsmaßnahmen, da Schulungsmaßnahmen unabhängig von variierenden Teilnehmerzahlen, einschließlich einer hohen Anzahl von Teilnehmenden, effizient bereitgestellt und durchgeführt werden können und
- die Kosteneffizienz, da sich die Kosten i.V. zu traditionellen Schulungsmethoden hinsichtlich z. B. der Reise- und Druckkosten reduzieren. [5]

Mit der cloudbasierten Bereitstellung eines LMS im Rahmen eines Software-as-a-Service-Modells (SaaS) erweitern sich diese Vorteile zusätzlich für Unternehmen, insbesondere durch u. a.:

- eine erhöhte Flexibilität und Erreichbarkeit, da der Zugriff auf ein LMS von jedem webfähigen Gerät aus möglich ist (mobiles Lernen);
- eine einfache Skalierbarkeit, da sich der Ressourcenbedarf leicht an variierende Teilnehmerzahlen anpassen kann und
- weitere Kostenreduktionen hinsichtlich der Infrastruktur und dem Betrieb.

Zielsetzung

Die zentrale Forschungsfrage dieser Arbeit lautet: *„Wie kann die Porsche AG Cloud Computing nutzen, um das unternehmensweite LMS bestmöglich zu optimieren?“*
Das Ziel der Arbeit ist daher die Entwicklung einer fundierten Handlungsempfehlung zur Auswahl eines cloudbasierten LMS-Anbieters, der die unternehmensspezifischen Anforderungen an ein LMS optimal erfüllt. Ergänzend werden weitere erforderliche Maßnahmen beschrieben, wie z. B. die genaue Überprüfung der Einhaltung unternehmensspezifischer IT-Sicherheitsanforderungen, um einen umfassenden und ganzheitlichen Ansatz für die Handlungsempfehlung sicherzustellen.
Ein wichtiges Zwischenziel besteht außerdem in der Ermittlung und Analyse der Anforderungen der verschiedenen Stakeholdergruppen, um die spezifischen Bedürfnisse der Hauptbeteiligten im Unternehmen bestmöglich bei der Anbieterauswahl zu berücksichtigen.

Vorgehensweise

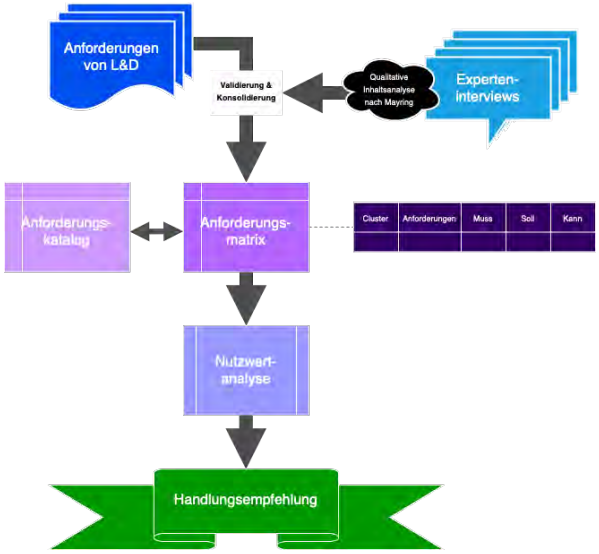


Abb. 2: Ablauf des Vorgehens [2]

Um die Forschungsfrage zu beantworten, wird ein mehrstufiges methodisches Vorgehen gewählt (siehe Abbildung 2). Zunächst wird eine umfassende Literaturrecherche durchgeführt, um die theoretischen Grundlagen zu schaffen und den aktuellen Stand sowie Entwicklungen im Bereich von Lernmanagementsystemen, Cloud Computing sowie deren Kombination zu erfassen. Parallel dazu werden bisher dokumentierte Anforderungen an ein LMS sowie weitere relevante Informationen von zuständigen Fachbereichen, insbesondere dem Learning & Development-Bereich (L&D), eingeholt. Die vorliegenden Informationen unterschiedlicher Art werden anschließend analysiert und tabellarisch konsolidiert, um eine strukturierte Übersicht über die bisherigen Anforderungen zu schaffen.
Für eine ganzheitliche Betrachtung der Anforderungen und Perspektiven verschiedener Stakeholdergruppen sowie zur Validierung der bisherigen Anforderungen, werden im darauffolgenden Schritt zwölf Experteninterviews durchgeführt. Die Auswahl der Experten orientiert sich dabei an den zentralen Benutzerrollen im aktuellen LMS, um die relevantesten Anforderungen für die Auswahl eines neues LMS systematisch zu erfassen. Darüber hinaus dient die Durchführung der Interviews dazu, Bewertungen des aktuellen Systems sowie mögliche Verbesserungspotenziale zu erheben. Der Interviewleitfaden umfasst insgesamt 36 Fragen, wobei nicht alle dieser Fragen an jeden Experten gestellt werden, sondern nur eine Auswahl der Fragen, die sich an der jeweiligen Systemrolle des Experten orientiert.

Es werden folgende Systemrollen aus dem aktuellen LMS befragt (siehe Abbildung 3):

- Führungskraft - umfasst insbesondere die Steuerung und Überwachung von Lernprozessen der Mitarbeitenden.
- Systemadmin - umfasst insbesondere die fachliche und technische Verwaltung des LMS.
- Systemadmin KoGe (Konzerngesellschaft) - umfasst die Verwaltung der entsprechenden Konzerngesellschaft.
- Qualifizierer - umfasst insbesondere die Erstellung und Verwaltung von Schulungsmaßnahmen.
- Lerner - umfasst insbesondere das Absolvieren von Schulungsmaßnahmen.

Abb. 3: Rollenverteilung der Experten im aktuellen LMS [2]

Die Interviewergebnisse werden mithilfe der Qualitativen Inhaltsanalyse nach Mayring analysiert und mit den bereits vorhandenen Anforderungen validiert und anschließend konsolidiert. Die finalen Anforderungen werden grundlegend in funktionale und technische Anforderungen unterteilt und in einer Anforderungsmatrix aufbereitet, die neben einer gesamthaften Visualisierung, eine kurze Beschreibung, das zugeordnete Cluster und eine Priorisierung der jeweiligen Anforderung umfasst. Ergänzt wird die Anforderungsmatrix durch einen Anforderungskatalog, der detailliertere Informationen zu den jeweiligen Anforderungen bereitstellt. Im nächsten Schritt wird eine Nutzwertanalyse basierend auf der Anforderungsmatrix erstellt und durchgeführt, mithilfe dessen mögliche Anbieter bewertet werden. Abschließend wird eine Handlungsempfehlung formuliert, die sowohl die Auswahl bestgeeigneter Anbieter als auch mögliche nächste Schritte umfasst.

Ausblick

Eine wesentliche Herausforderung besteht darin, eine Standardlösung zu implementieren, die zugleich genügend Flexibilität für unternehmensspezifische Anpassungen bietet. Vor dem Hintergrund eines großen Unternehmens mit zahlreichen abzubildenden Use Cases, die auch die bisherige (customized) LMS-Lösung umfangreich unterstützt, gewinnt diese Herausforderung zusätzlich an Komplexität. Das Ziel besteht darin, die Vorteile von Standardlösungen vollständig zu bewahren und jedoch gleichzeitig das System so anpassen zu können, dass es die wichtigsten unternehmensspezifischen Anforderungen erfüllt.
Langfristig ist es außerdem essenziell, die dynamischen Entwicklungen im Bereich Corporate (E-)Learning kontinuierlich zu beobachten und zu berücksichtigen. Aktuelle Trends wie der Übergang zu skill-driven Learning und der Einsatz von Lernökosystemen zeigen, dass die Anforderungen an LMS nicht statisch sind, sondern stetig weiterentwickelt werden müssen.
Abschließend bleibt es entscheidend, ein LMS zu wählen, das nicht nur den gegenwärtigen Anforderungen gerecht wird, sondern auch langfristig skalierbar und anpassungsfähig ist, um den Wandel im Corporate-Learning nachhaltig zu unterstützen

Literatur und Abbildungen

[1] Alexandra Abletshauser. Was ist ein Learning Management System (LMS)? https://www.haufe-akademie.de/digital-suite/blog/lms-was-ist-ein-learning-management-system, 2024.

[2] Eigene Darstellung.

[3] Information Technology Gartner. Definition of E-Learning. https://www.gartner.com/en/information-technology/glossary/e-learning, 2024.

[4] Ingrid Hovdar-Stojakovic et al. *Innovatives Lehren und Lernen mit Blended Learning*. Springer Fachmedien, 2023.

[5] Frank Siepmann. *Teilstudie: Digitales Lernen in der betrieblichen Bildung*. eLearning Journal, 2024.

[6] Stefan Strohmeier. *Informationssysteme im Personalmanagement*. Vieweg+Teubner, 2008.

Datenerfassung und Auswertung von Nutzerinteraktionen in Bedienoberflächen von Fertigungsmaschinen für die Produktoptimierung - Machbarkeitsnachweis am Beispiel einer webbasierten HMI von Bosch Connected Industry

Maike Knauer **Dirk Hesse**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Bosch Connected Industry, Stuttgart-Feuerbach

Einleitung

Die Auswertung der eigenen Webseite, sowohl von Privatpersonen als auch der Webauftritt eines Unternehmens, kann heutzutage mithilfe von Web Analytics Tools wie beispielsweise Google Analytics ohne großen Aufwand durchgeführt werden. Bereits nach wenigen Klicks lassen sich erste Daten wie Pageviews, Scrolls und weitere Events, die auf der Webseite stattfinden, auslesen, in einem ansprechenden Dashboard darstellen und Schlüsse daraus ziehen, an welchen Stellen die Webseite noch Verbesserungspotenzial hat.
Weniger simpel gestaltet sich das Vorgehen, wenn es sich nicht um eine Webseite mit statischen Inhalten handelt, die analysiert und ausgewertet werden soll, sondern um eine Webapplikation mit dynamischen Inhalten, die als Mensch-Maschine-Schnittstelle einer Fertigungsmaschine zur Bedienung der Maschine dient und keine Internetverbindung hat. Die folgende Arbeit zeigt auf, wie man eine solche Webapplikation analysieren kann.

Zielsetzung der Arbeit

Ziel dieser Arbeit ist die Produktoptimierung der Bedienoberfläche von Fertigungsmaschinen von Bosch Connected Industry (BCI). Dafür soll einerseits ermittelt werden, welche Ansichten und Buttons von Nutzern besonders oft verwendet werden, sodass bei der weiteren Entwicklung ein Fokus auf die Optimierung der Elemente gelegt werden kann. Außerdem soll durch das Aufzeichnen der verschiedenen Events wie Klicks oder Sichtbarkeitsänderungen der Pfad von Benutzern in der Mensch-Maschine-Schnittstelle verfolgt werden, um im Zuge der Produktoptimierung beispielsweise Wege zu vereinfachen.
Das übergeordnete Ziel besteht darin, die technische Machbarkeit der zuvor genannten Ziele nachzuweisen. Aufgrund dessen wird diese Arbeit als Proof of Concept durchgeführt wobei rechtliche Fragestellungen, wie beispielsweise die Zulässigkeit der Erhebung und Speicherung personenbezogener Daten, ausgeklammert werden.

Forschungsfragen

Um diese Ziele zu erreichen und die zuvor beschriebene Problemstellung zu lösen, wird die Forschungsfrage „Wie kann eine webbasierte Bedienoberfläche einer Fertigungsmaschine basierend auf Daten aus Product Analytics optimiert werden?“ formuliert und in drei Teilfragen untergliedert.

1. Welche Daten von Nutzerinteraktionen mit der Web-HMI einer Fertigungsmaschine können ohne Internetverbindung erfasst werden und welche Herangehensweise bietet sich für die Datenerfassung an?
2. Welche Art der Analyse kann für die Auswertung genutzt werden und welche Tools können gegebenenfalls zur Unterstützung verwendet werden?
3. Wie können aus den gesammelten und analysierten Daten Schlüsse für die Produktoptimierung gezogen werden?

Die Antworten auf diese Fragen sollen im Rahmen der Masterarbeit erarbeitet werden.

Mensch-Maschine Interaktion

Butz und Krüger beschreiben in [2], dass sich aufgrund der Funktionsweise vieler Alltagsgeräte mittels Computertechnologie zwischen einem Computer und einer Maschine meist keine klare Grenze mehr ziehen lässt. Aufgrund dessen wird argumentiert, dass die

Begriffe „Mensch-Maschine Interaktion“ und „Mensch-Computer-Interaktion“ synonym verwendet werden können.
Die Interaktion zwischen Mensch und Maschine erfolgt über eine Mensch-Maschine-Schnittstelle (engl. Human machine interface, HMI). Die HMI verfügt meist über eine grafische Oberfläche und besteht aus zwei Elementen, einerseits aus der Anzeige der Parameter, des Status sowie der Prozesse der Maschine und andererseits aus der Steuerung, in der ein Mensch Eingaben tätigt, die anschließend von der Maschine ausgeführt werden. [7]

Interaktionsgeräte

Dahm gibt in [3] einen Überblick über die Interaktionsgeräte, die es Anwendern ermöglichen, mit der Maschine zu interagieren und die HMI zu verwenden. Das sind:

- "Tastaturen[...]
- Zeigegeräte wie Maus oder Trackball mit den typischen Aktionen
- Touchscreen, der Eingabe und Ausgabe in einem Gerät vereinigt
- Natürliche Sprache für Ein- und Ausgabe
- Befehlssprachen zur Steuerung“ [3]

Während beispielsweise mit der Maus der genaue Verlauf des Zeigers auf dem Bildschirm nachverfolgt werden kann, muss beim Touchscreen auf die "Berührung [des Bildschirms] mit einem (Single-Touch) oder mehreren Fingern oder Handflächen (Multi-Touch)"[2] gewartet werden.

HMInow

Im Rahmen dieser Arbeit wird die Bedienoberfläche für Fertigungsmaschinen namens HMInow von Bosch Connected Industry, wie sie in Abbildung 1 dargestellt ist, analysiert. HMInow ist eine webbasierte und vom Hersteller der speicherprogrammierbaren Steuerung unabhängige HMI-Lösung. [5] Mithilfe eines Baukastens, der verschiedene Controls enthält, können "schnell einheitliche und nutzerfreundliche Bedienoberflächen für Bediener [...] von Maschinen“ [5] entwickelt werden. Sie basiert auf der VisiWin Web UI von Inosoft. [6] Die HMInow ist mit den Web-Technologien HTML und JavaScript gebaut und läuft auf der V8 Engine in Chrome. Für den Betrieb benötigt die HMInow zwar keine Internetverbindung, jedoch muss eine Netzwerkverbindung zum VisiWin Webserver bestehen. Über die Netzwerkverbindung erhält die HMInow Informationen vom Webserver und wird zur Laufzeit dynamisch aufgebaut.

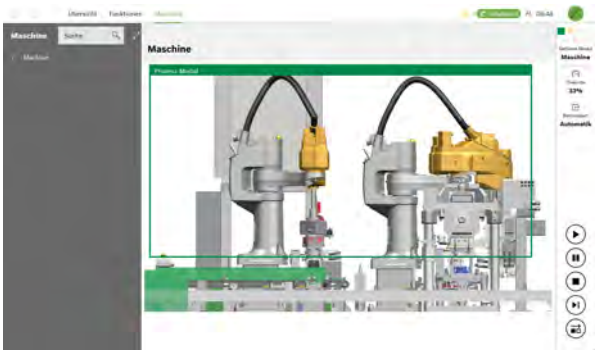


Abb. 1: HMInow [4]

Im Rahmen dieser Arbeit wird von der Annahme ausgegangen, dass die Bedienoberfläche einer Fertigungsmaschine lediglich aus einem Touchscreen zur Anzeige der Web-HMI besteht.

Product Analytics

Product Analytics hat seinen Ursprung in den Bereichen der Business Intelligence sowie der Web-Analytics. [11] Allerdings geht Product Analytics über die reine Datensammlung hinaus und analysiert systematisch, häufig eventbasiert, Produkte und Nutzer von Websites, mobilen Apps oder Social-Media-Plattformen. Aus dieser Analyse sollen anschließend wichtige Erkenntnisse gewonnen werden, mit deren Hilfe Produkte und Dienstleistungen verbessert und die Nutzerzufriedenheit gesteigert werden kann.
Die aktuelle, weltweite Relevanz der Beschäftigung mit Product Analytics wird durch eine Prognose des Marktforschungsinstituts Polaris, siehe Abbildung 2, bestätigt. Auf Basis historischer Daten von 2018 bis 2020 prognostiziert Polaris eine Wachstumsrate von 21,1% für den Product Analytics Markt bis zum Jahr 2030. [10]

Abb. 2: Product Analytics Market Size Prediction [10]

Witzenleiter identifiziert in [11] Mixpanel, Amplitude, Posthog, Pendo, Heap und Fullstory als die gängigsten Lösungen und beliebtesten Tools für Product Analytics. Da für die Analyse der HMInow jedoch keine Inter-

netverbindung besteht, kommen nur selbst gehostete Lösungen in Frage, die im lokalen Netz gehostet werden können und somit keine Internetverbindung benötigen. Dies trifft von den genannten Tools lediglich auf Posthog und Amplitude zu, dementsprechend werden die beiden Tools für den weiteren Einsatz in Betracht gezogen.

Verwandte Arbeiten

Maxwell und Hauff haben im Jahr 2021 mit LogUI ein Open-Source Logging Framework entwickelt, das laut eigener Aussage in der Lage ist „praktisch jede Benutzerinteraktion auf einer Webseite“ [8] zu erfassen. Da das Framework frei verfügbar ist und die Anforderungen für einen Betrieb im lokalen Netzwerk erfüllt, sind alle erforderlichen Krieterien an ein Product Analytics Framework gegeben, sodass im Rahmen dieser Arbeit untersucht, ob das LogUI Framwork auch den Anforderungen zur Datensammlung entspricht und eingesetzt werden kann.
Im Jahr zuvor haben Solís-Martínez et al. UXJs entwickelt. [9] Sie beschreiben UXJs als neuartigen Forschungsansatz zur automatischen Erfassung aller möglichen Informationen über die Nutzeraktivität auf Websites, der diese Informationen quantitativ darstellt und ihre automatische statistische Analyse und das schnelle Verständnis durch Webentwickler ermöglicht. Da UXJs jedoch nicht veröffentlicht ist, kann dieses Tool nicht weiter in Betracht gezogen werden.
Bereits vor zehn Jahren hat sich eine Masterarbeit damit beschäftigt, „wie Software-Analytik zur Sammlung von Daten über die Nutzung von Web- und Mobilanwendungen eingesetzt werden kann und wie die mit Analytics gesammelten Daten zur Verbesserung dieser Arten von Anwendungen beitragen können“. [1] Da dies ein Teil der angestrebten Arbeit dieser Thesis ist, muss untersucht werden, ob sich die Vorgehensweise von Ahola möglicherweise ganz oder in Teilen auf diese Arbeit übertragen lässt.

Anforderungen zur Datensammlung

Eine Anforderung an die Implementierung der Datensammlung besteht darin, dass ein Wechsel zwischen verschiedenen Ansichten der HMI erkannt werden muss auch wenn sich die URL nicht ändert, da die HMInow eine Single-Page-Webanwendung ist.
Außerdem darf die Datensammlung nur mithilfe eines Tools durchgeführt werden, falls dieses nicht auf einer Cloud des Anbieters gehostet wird, sondern lokal gehostet werden kann, da es sich bei den Daten der HMInow um intern klassifizierte Unternehmensdaten handelt.
Des Weiteren müssen Daten zu den von Usern verwendeten Ansichten und Controls gesammelt werden können. Auch die Reihenfolge der verwendeten Ansichten und Controls muss erfasst werden, um anschließend Auswertungen vornehmen zu können.

Planung der Umsetzung

Die Auslassung der Betrachtung rechtlicher Fragestellungen fürhrt dazu, dass keine realen Nutzerinteraktionsdaten gesammelt werden können. Dies begründet sich darin, dass Benutzerdaten nicht ohne weiteres gespeichert und analysiert werden dürfen. Aufgrund dessen kann der Machbarkeitsnachweis nur in einem Testaufbau erprobt werden. Dafür werden zunächst verschiedene Use Cases formuliert. Diese sollen dann durch Software-Applikateure an der Teststation durchgeführt werden.

Ausblick

Im weiteren Verlauf der Arbeit soll die beschriebene Methodik umgesetzt werden, sodass zunächst Nutzerinteraktionsdaten der HMInow gesammelt und anschließend ausgewertet werden können. Zuvor muss jedoch entschieden werden, welches Product Analytics Tool zur Durchführung der Analyse in die HMInow integriert werden soll. Anschließend folgt die Implementierung der Datenübertragung vom Client an den Server, sodass daraufhin die Auswertung der Nutzerinteraktionsdaten starten kann. Dafür müssen mögliche Auswertungsmethoden auf Eignung für diesen Anwendungsfall evaluiert werden.

Literatur und Abbildungen

[1] Juha Ahola. Designing with Data: Using Analytics to Improve Web and Mobile Applications, 2014.

[2] Andreas Butz and Antonio Krüger. *Mensch-Maschine-Interaktion*. De Gruyter Oldenbourg, 2014.

[3] Markus Dahm. *Grundlagen der Mensch-Computer-Interaktion*. Pearson-Studium, ein Imprint von Pearson Education, 2006.

[4] Eigene Darstellung.

[5] Andreas Deininger. HMInow. https://inside-docupedia.bosch.com/confluence/display/NAPD/HMInow, 2024.

[6] Inosoft GmbH. Plattformunabhängige Prozessvisualisierung im Browser mit VisiWin Web UI. https://www.inosoft.com/produkt/mobile-web-hmi/, 2024.

[7] Naveen Kumar and Seul Chan Lee. Human-machine interface in smart factory: A systematic literature review. *Technological Forecasting and Social Change*, 2022.

[8] David Maxwell and Claudia Hauff. LogUI: Contemporary Logging Infrastructure for Web-Based Experiments. In *Advances in Information Retrieval*, pages 525–530. Springer International Publishing, 2021.

[9] Jaime Solís-Martínez et al. UXJs: Tracking and Analyzing Web Usage Information With a Javascript Oriented Approach. *IEEE Access*, pages 43725–43735, 2020.

[10] Polaris Market Research und Consulting Inc. Product Analytics Market Share, Size, Trends, Industry Analysis Report, By Component (Solutions, Services), By Deployment (Cloud, On-premises), By Industry Vertical; By Region; Segment Forecast, 2022 - 2030. https://www.polarismarketresearch.com/industry-analysis/product-analytics-market, 2024.

[11] Michael Witzenleiter. *Quick Guide Product Analytics: Wie Sie mit Systemen wie Google Analytics 4 und Co. mehr über Ihre Nutzer und deren Produktakzeptanz lernen können*. Springer Gabler, 2023.

Video-Deepfakes im Fokus: Eine vergleichende Analyse der Effektivität von Open-Source-Tools bei der Identifizierung und Generierung manipulierter Videos prominenter deutscher Persönlichkeiten

Glykeria Koutsianou **Dieter Morgenroth**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Motivation und Problemstellung

Die Verbreitung manipulativer Inhalte, welche durch Deepfake-Technologien erzeugt werden, stellt eine zunehmend ernstzunehmende Bedrohung für die Authentizität öffentlicher Kommunikation dar. Insbesondere Videos, welche prominente Persönlichkeiten in fiktiven Situationen darstellen, bergen das Potenzial, das Vertrauen in Medien, politische Prozesse und öffentliche Institutionen zu untergraben. Die Verbreitung derartiger Inhalte kann zur Förderung von Desinformationskampagnen, zur Erschütterung des Vertrauens in Institutionen sowie zu gravierenden Schäden führen.

In der deutschen Forschungslandschaft besteht bislang ein Mangel an spezifischen Untersuchungen zur Effektivität von Erkennungswerkzeugen gegen derartige gezielte Manipulationen. Die kontinuierliche Optimierung von Deepfake-Generatoren erfolgt in vielen Fällen schneller als die Entwicklung von Erkennungsmethoden und Datensätzen, die als Grundlage dienen. Dies führt zu einer Beeinträchtigung der Zuverlässigkeit der Tools. In diesem Kontext ist zu hinterfragen, inwiefern die gegenwärtig verfügbaren Erkennungsmethoden den Anforderungen realer Bedrohungsszenarien tatsächlich gerecht werden und wie zuverlässig die bestehenden Tools sind.

Die vorliegende Bachelorarbeit untersucht die Effektivität aktueller Erkennungstools bei der Analyse gezielt generierter Deepfakes, die deutsche Persönlichkeiten des öffentlichen Lebens betreffen. Ziel ist die Identifikation von Schwachstellen und Potenzialen der Tools, um technische Handlungsempfehlungen abzuleiten für die Notwendigkeit Datensätze prominenter deutscher Persönlichkeiten. Die Ergebnisse sollen dazu beitragen, die Sicherheit im digitalen Informationsraum zu stärken und das Vertrauen in digitale Inhalte langfristig zu sichern.

Vorgehen

Die Effektivität von Deepfake-Erkennungstools wird im Rahmen eines experimentellen Ansatzes untersucht, der sowohl technologische Bewertungen als auch menschliche Einschätzungen kombiniert. Im Rahmen der Untersuchung wird zunächst das Open-Source-Tool Face-Fusion [6] eingesetzt, um gezielt 30 Video-Deepfakes zu erstellen. Dabei basieren 15 der erstellten Deepfakes auf prominenten deutschen Persönlichkeiten, während die übrigen 15 Gesichter von nicht prominenten Personen darstellen, bespielhaft in Abbildung 1 zu sehen. Die Aufteilung erlaubt eine vergleichende Analyse der Erkennungseffizienz in Bezug auf den Bekanntheitsgrad der Zielpersonen. Aufgrund ihrer medialen Präsenz sind prominente Gesichter in Datensätzen häufiger vertreten, wodurch sie potenziell leichter zu erkennen sind [3] [2].

Abb. 1: Original Video und Deepfake nebeneinander gestellt [1]

Zur technischen Evaluierung wird das Deepfake-Erkennungswerkzeug "Deepfake-O-Meter"[5] eingesetzt, welches auf maschinellem Lernen basiert und gezielt für die Analyse manipulierter Videoinhalte entwickelt wurde. Das Tool liefert quantitative Ergebnisse hinsichtlich der Wahrscheinlichkeit, dass ein Video ein Deepfake ist. Dabei kann aus verschieden trainierten Erkennungsmodellen gewählt werden, welche sich

durch ihren Datensatz und Erkennungsmethode unterscheiden. Dies erlaubt eine systematische Evaluierung der Tools hinsichtlich ihrer Erkennungsgenauigkeit in beiden Szenarien (prominent vs. nicht prominent).
In Ergänzung dazu erfolgt die Präsentation der generierten Videos in einer Umfrage, um subjektive Wahrnehmungen und Einschätzungen von Probanden zu erfassen. Die Teilnehmerinnen und Teilnehmer werden gebeten, eine Einschätzung darüber abzugeben, ob sie die präsentierten Videos als authentisch oder manipuliert einstufen. Dies erlaubt wertvolle Rückschlüsse auf die Grenzen der menschlichen Erkennung im Vergleich zu technischen Tools [4]. Das gesamte Vorgehen wird in Abbildung 2 visualisiert.

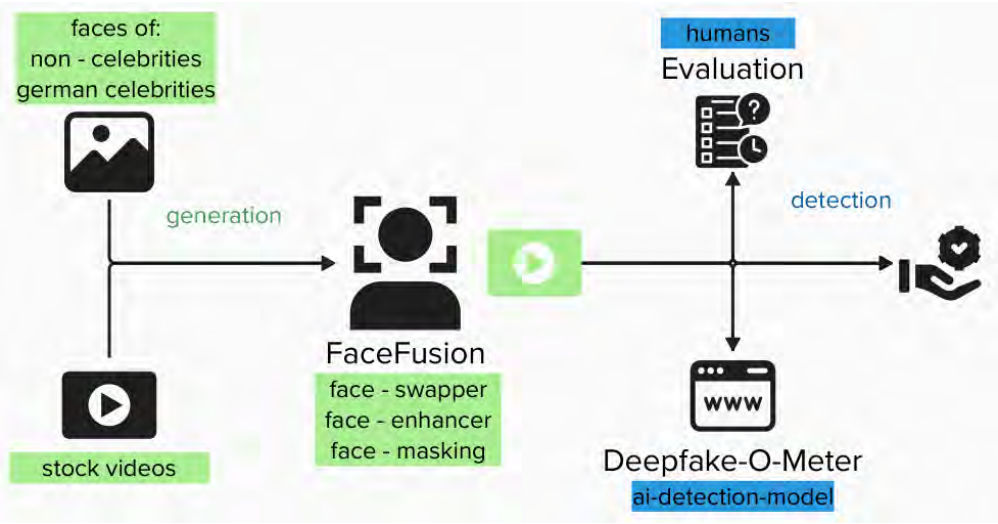


Abb. 2: Pipeline der Untersuchung der Bachelorarbeit [1]

Die Kombination aus technischer Analyse und menschlicher Einschätzung erlaubt eine fundierte Einschätzung der aktuellen Effektivität von Deepfake-Erkennungstools. Gleichzeitig werden potenzielle Schwächen bei der Bewertung gezielt generierter Inhalte aufgedeckt, insbesondere im Kontext von Prominenten, die aufgrund ihrer medialen Präsenz oft im Fokus solcher Technologien stehen.

Ausblick und Zusammenfassung

Die Bachelorarbeit demonstrierte die Video-Generierung und verdeutlichte, dass diese Technologien inzwischen leicht zugänglich und leistungsstark sind. Die Fähigkeit zur Erstellung täuschend echter Deepfakes demonstriert nicht nur die Fortschritte im Bereich der KI, sondern auch die mit ihrer breiten Verfügbarkeit einhergehenden Risiken. Die Analyse der generierten Videos offenbarte sowohl Stärken als auch Schwächen des eingesetzten Erkennungstools. Die Analyse der generierten Videos hat ergeben, dass deutsche prominente Gesichter von den Tools nicht mit einer höheren Trefferquote erkannt wurden. Bei nicht prominenten Gesichtern war die Erkennungsrate geringer, was die Notwendigkeit für umfassendere und diversere Ansätze in der Entwicklung von Detektionsmethoden unterstreicht. Die Resultate liefern wichtige Erkenntnisse über den aktuellen Stand der Deepfake-Erkennung. Sie verdeutlichen den technologischen Fortschritt, aber auch das Gefahrenpotenzial gezielter Deepfakes, die bestehende Algorithmen umgehen können. Zukünftige Forschung sollte sich daher auf die Entwicklung robusterer und anpassungsfähigerer Erkennungswerkzeuge mit erweiterten Datensätzen konzentrieren.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] David Güera. Deepfake Video Detection Using Recurrent Neural Networks. https://doi.org/10.1109/AVSS.2018.8639163, 2018.

[3] Pavel Korshunov. DeepFakes: a New Threat to Face Recognition? Assessment and Detection. https://doi.org/10.48550/arXiv.1812.08685, 2018.

[4] Yuezun Li. In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking. https://doi.org/10.1109/WIFS.2018.8630787, 2018.

[5] Siwei Lyu. Deepfake-O-Meter. https://zinc.cse.buffalo.edu/ubmdfl/deep-o-meter/, 2024.

[6] Henry Ruhs. FaceFusion. https://github.com/facefusion/facefusion, 2024.

Konzeption und Realisierung von Docs-as-Code-Toolstacks zur automatisierten Generierung technischer Dokumentationen

Damaris Kroener **Harald Melcher**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma pep.digitial GmbH, Esslingen (Neckar)

Softwaredokumentation ist ein essentieller Bestandteil der Softwareentwicklung. Sowohl Endnutzerinnen und Endnutzer, als auch Entwicklerinnen und Entwickler ziehen sie zu Rate, um die Software zu nutzen, zu warten oder weiterzuentwickeln. Dies erfordert allerdings, dass die Dokumentation aktuell und vollständig ist. Das hängt zu großen Teilen von den Entwicklerinnen und Entwicklern der Software ab, diese sehen sich jedoch mit einigen Hürden konfrontiert (z.B. Medienbrüche, Binäre Dateiformate, fehlende Versionierung), die das Dokumentieren erschweren.

Dieser Problematik soll die Philosophie „Docs-as-Code“ (kurz für „treat documentation as code“) entgegenwirken. Diese hat das Ziel „Dokumentation in einem Entwicklungsvorhaben genauso zu behandeln wie den Quelltext“ [4]. Das bedeutet, dass Entwicklerinnen und Entwickler die Dokumentation mit denselben Werkzeugen schreiben wie den Quellcode ihrer Software.

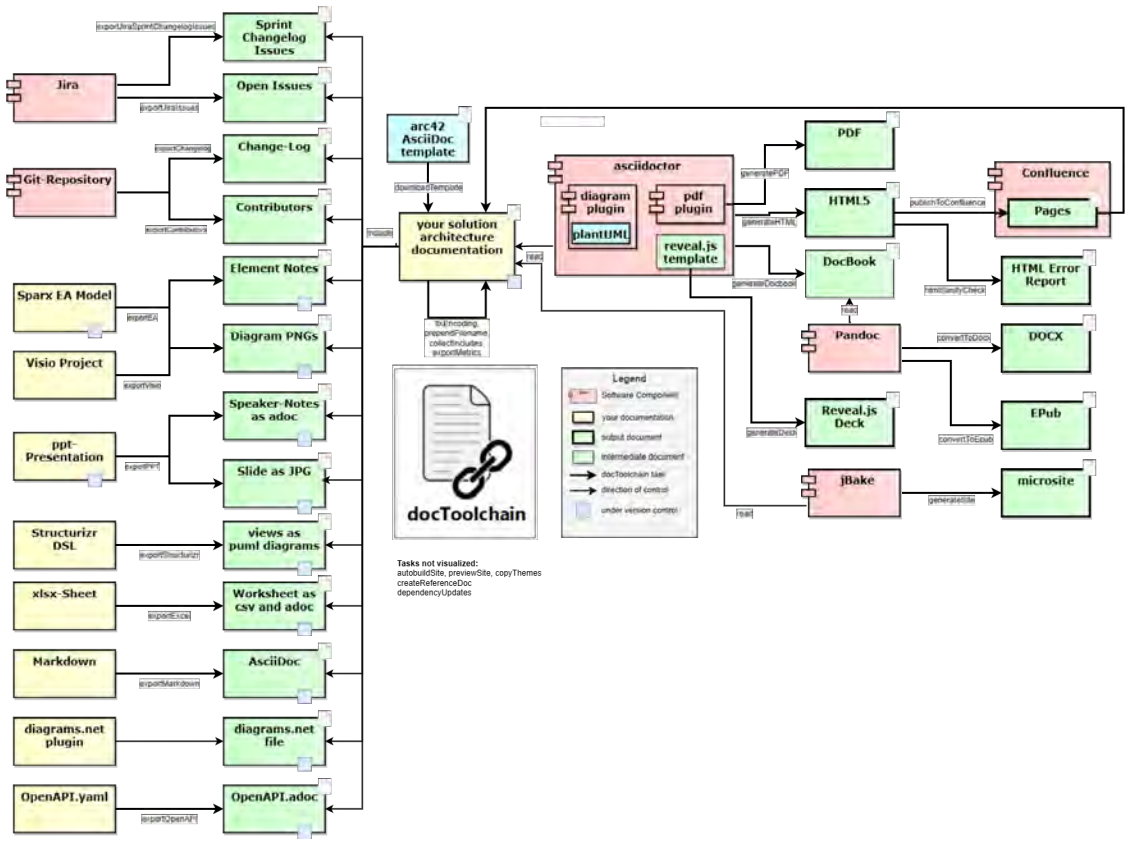


Abb. 1: Taskübersicht von docToolchain [2]

Um dessen Umsetzung zu erleichtern, können Entwickler das Open-Source Projekt „docToolchain“ verwenden. Dabei handelt es sich um eine Sammlung von Skripten, die die Umsetzung von Docs-as-Code erleichtern. docToolchain ist ein Open-Source Projekt, das von Ralf D. Müller mit dem Ziel gegründet wurde, das Erstellen und Pflegen technischer Dokumentation nach der Docs-as-Code Philosophie zu erleichtern. docToolchain verwendet die Aufzeichnungssprache AsciiDoc. [3]
Mit ihren Features (genannt „Tasks“, da docToolchain auf gradle basiert) unterstützt docToolchain unter anderem das Rendern von Quelldateien der Dokumentation in verschiedenen Formaten (z.B. HTML, PDF), das Exportieren von Bildern und AsciiDoc-Schnipseln aus anderen Systemen oder Dateiformaten (z.B. Jira, Excel), das Konvertieren von AsciiDoc-Dateien in andere Formate (z.B. EPUB, docx) und das veröffentlichen von Dokumentation zu Plattformen wie Confluence 1.
Dieser Toolchain fehlen allerdings noch einige Features. Besonders Tasks, die tatsächlich Informationen aus dem Sourcecode herausziehen und diese in eine Form bringen, die direkt in eine Dokumentation eingebunden werden kann, sind kaum vorhanden. Aus diesem Grund erforscht der praktische Teil der Bachelorarbeit, wie docToolchain um neue Features erweitert werden kann. Das Open-Source Projekt „Spring PetClinic“ dient dafür als Beispielprojekt, aus dem drei neuen Tasks in unterschiedlicher Weise Inhalte extrahieren sollen. Die Ergebnisse dieser Tasks werden in ein arc42-Template eingefügt und mithilfe von bereits existierenden docToolchain-Tasks auf einem Confluence-Space veröffentlicht.

1.3.1. Git Contributions

Name	Email	Commit Count
Anna Otto	anna-otto@email.com	3
John Doe	john@doe.org	3
Max Mustermann	max.muster@mail.de	2

Abb. 2: Ergebnis des Tasks exportGitContributions [1]

Der erste entwickelte Task „exportGitContributions“ beschäftigt sich mit der tabellarischen Darstellung der Mitwirkenden an einem Git-Repository und der Anzahl ihrer Commits 2. Die Problematik in der Umsetzung besteht darin, dem Task explizit zu übergeben, in welchem Verzeichnis er Kommandozeilenbefehle ausführen soll, da Befehle wie „git shortlog“ nur innerhalb eines Projektes mit Git funktionieren.

13.33. angular

Publisher/Author/Vendor	Angular Authors		
Group	@schematics		
Version	16.2.0		
Description	Schematics specific to Angular		
Licenses	MIT		
Package URL	pkg:npm/%40schematics/angular@16.2.0		
External References	git+https://github.com/angular/angular-cli.git	vcs	as detected from PackageJson property "repository.url"
	https://github.com/angular/angular-cli	website	as detected from PackageJson property "homepage"
	https://github.com/angular/angular-cli/issues	issue-tracker	as detected from PackageJson property "bugs.url"
	https://registry.npmjs.org/@schematics/angular/-/angular-16.2.0.tgz	distribution	as detected from npm-ls property "resolved" and property "integrity"
Properties	N/A	node_modules/@schematics/angular	
	N/A	true	
Components	N/A		N/A

Abb. 3: Tabellarische Darstellung des SBOMs für die Komponente angular [1]

„exportSBOM“ befasst sich mit der gesetzlich vorgeschriebenen Notwendigkeit (NIS2, CRA) für ein Softwareprodukt Informationen und Schwachstellen mithilfe von SBOMs („Software Bill of Materials“) offenzulegen. Dieser Task verwendet dafür die CycloneDX Plugins, um aus dem PetClinic Projekt SBOMs zu generieren. Die sich daraus ergebenen JSON-Dateien konvertiert der Task anschließend in Tabellen. Jede Komponente erhält dabei ihre eigene Tabelle 3.
Der Task „exportCodeSnippets“ ist der vorerst letzte geplante Task. Dieser Task soll dazu Autorinnen und Autoren der Softwaredokumentation dabei helfen, Codestellen aus dem Sourcecode in der Dokumentation referenzieren zu können. Um sicherzustellen, dass diese Codeausschnitte stets auf dem neuesten Stand sind, soll der Task bei jedem Bau der Dokumentation die aktuelle Version der Codestelle einbinden. Zum Zeitpunkt des Schreibens dieses Artikels befindet sich dieser Task noch in Arbeit.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Ralf .D Müller. What Is a Task? https://doctoolchain.org/docToolchain/v2.0.x/015_tasks/03_tasks.html, 2024.

[3] Gernot Starke and Peter Hruschka. *arc42 in Aktion: praktische Tipps zur Architekturdokumentation*. Hanser, 2023.

[4] Stefan Zörner. *Software-Architekturen dokumentieren und kommunizieren: Entwürfe, Entscheidungen und Lösungen nachvollziehbar und wirkungsvoll festhalten*. Carl Hanser Verlag GmbH & Co. KG, 2022.

Effiziente Datenanalyse in der Fahrzeugsicherheit: Automatisierte Prüfung von Airbag-Auslösezeiten

Enes Kuecuekakyuez **Astrid Beck**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Robert Bosch GmbH, Bietigheim

Effiziente Datenanalyse in der Fahrzeugsicherheit: Automatisierte Prüfung von Airbag-Auslösezeiten

Einleitung zur passiven Sicherheit im Straßenverkehr

In den letzten Jahrzehnten hat die Sicherheit in der Automobilbranche große Fortschritte erzielt, wobei die Entwicklung von der Airbag Technologie eine erhebliche Rolle einnimmt. Zu Beginn der Automobilproduktion war die Ausstattung von Fahrzeugen mit Sicherheitssystemen noch nicht üblich. In der heutigen Zeit ist ein Fahrzeug ohne Sicherheitssysteme unvorstellbar. Die Airbags stellen eine essenzielle Komponente der passiven Sicherheitsausstattung dar, welche zusammen mit dem Sicherheitsgurt ein erfolgreiches Rückhaltesystem bilden. Ein Blick auf die zeitliche Entwicklung zeigt, dass durch Innovation und gesetzliche Regelungen signifikante Fortschritte erzielt wurden. Die Abbildung verdeutlicht den Rückgang der Verkehrstoten seit den 1950er Jahren in Deutschland und zeigt die Auswirkungen verschiedener gesetzlicher Maßnahmen. [2]

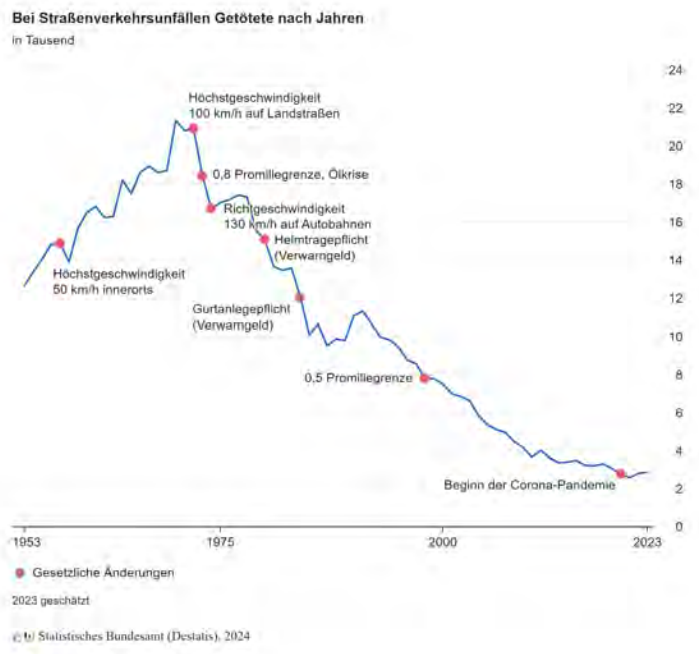


Abb. 1: Rückgang der Verkehrstoten seit den 1950er Jahren in Deutschland [4]

Neben dem Fortschritt physischer Sicherheitsmechanismen hat sich die rechnerische-Simulation als große Hilfe in der Fahrzeugentwicklung bewährt. Während physische Fahrzeug-tests helfen, Schwachstellen zu identifizieren, sind sie außerordentlich zeit- und kostenintensiv. Um diese Schwächen auszubessern, entwickelte sich die rechnerische Simulation zu einem anerkannten Entwicklungswerkzeug. Das

Simulieren reicht von der Konzeptphase bis hin zur Serienentwicklung aufgrund äußerst genauer und zuverlässiger Berechnung des Fahrzeugverhaltens und der Insassensicherheit. Dadurch werden nicht nur die Entwicklungszyklen verkürzt, es wird auch eine merkliche Verbesserung der Sicherheit im Fahrzeug erreicht. [3]

Ziel meiner Arbeit

Zielsetzung dieser Bachelorarbeit ist es, einen automatisierten Prozess zur Auswertung von Crash-Daten aus einer Datenbank zu entwickeln, welche alle Informationen zu den Crash-Simulationen enthält und in einer standardisierten Excel-Tabelle darzustellen. Dabei soll sichergestellt werden, dass die in den Daten enthaltenen Auslösezeiten vom Airbag zu den bestimmten Crashtypen mit den Vorgaben aus einem Requirement Sheet abgeglichen werden. Abweichungen zwischen den entnommenen Auslösezeiten und den Vorgaben sind zu identifizieren und visuell hervorzuheben, um Fehlauslösungen erkennen zu können.
Im Folgenden werden die benutzten Tools für die Fahrzeugsicherheit vorgestellt.

MdSng

Ein wichtigstes Kalibrierungstool MdSng ist für das Einstellen von Parametern für Datenbanken zuständig. Das Programm wird eingesetzt, um die Kalibrierungsparameter vom AIDA-Algorithmus für die Kunden anzupassen. Zudem dient es der Generierung von Dateien, die von anderen Kalibrierungstool verwendet werden können.

RSDBnext

Das Tool RSDBnext dient der Umwandlung von Daten aus einer Datenbank in ein lesbares Format, bei diesem als Zielformate die Dateiformate XML und XLSX vorgesehen sind. Es wird vor allem für das Extrahieren und Verarbeiten von Crash Daten nach dem Simulieren genutzt. Die Datenbank enthält Ergebnisse aus der Simulation, die in mehreren Tabellen gespeichert werden. Das Auslesen der Ergebnisse erfolgt durch Queries, die die Daten aus den Tabellen entnehmen. Anschließend werden alle relevanten Daten übernommen und in eine XML-Datei geschrieben. Mithilfe eines Stylesheets erfolgt danach die Umwandlung in eine lesbare Excel-Datei. Bei der Ausführung von RSDBnext kann der User festlegen, in welchem Format die Daten ausgegeben werden sollen. Dabei stehen verschiedene Optionen zur Verfügung:

- Result file
- Scaling File
- Misuse Report
- Standard Crash Report
- Result File Plus

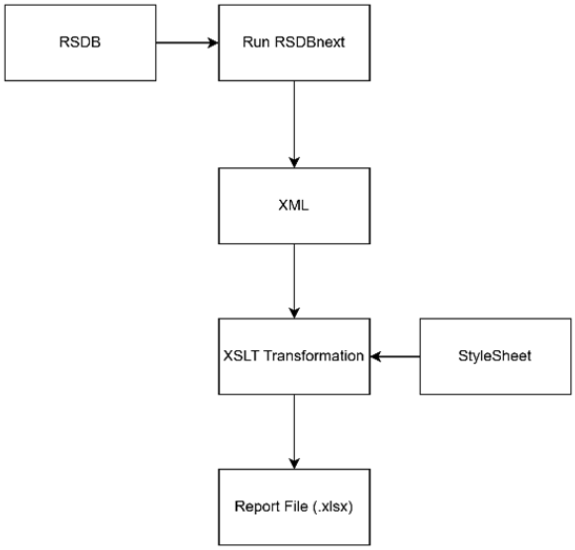


Abb. 2: Zeigt den Prozessablauf zur automatisierten Auswertung und Transformation von Crash-Daten, Der Ablauf beginnt in der Datenbank und endet mit der Erstellung eines standardisierten Berichts in Excel-Format. [1]

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Johannes Holtz. *Unfallverletzungen in Fahrzeugen mit Airbags*. Bremen : Fachverlag NW in Carl Ed. Schünemann KG, Februar 2022, 2022.

[3] Florian Kramer. *Passive Sicherheit von Kraftfahrzeugen : Biomechanik — Simulation — Sicherheit im Entwicklungsprozess*. Wiesbaden : Vieweg+Teubne, 3 edition, 2009.

[4] Bundesamt Statisches. Verkehrsunfälle. https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Verkehrsunfaelle/_inhalt.html, 10 2024.

Definition der Anforderungen an ein Leiterplattenlayout für IC-Level EMV-Tests nach IEC 62132-4 — Realisierung der Leiterplatte und Durchführung des Tests am Beispiel eines Schaltreglers

Lukas Kurz **Walter Lindermeir**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma MPS Germany GmbH, Ettenheim

Einleitung und Motivation

In der modernen Elektrotechnik stellt die elektromagnetische Verträglichkeit (EMV) einen zentralen Aspekt dar. Die ersten Anforderungen an integrierte Schaltkreise (ICs) bezüglich EMV wurden bereits 1965 gestellt. Seitdem wurden die Strukturen in den ICs kontinuierlich kleiner, von 10 µm in den 1970er Jahren auf heute unter 5 nm [6]. Die Funktionen, die in einem IC implementiert werden konnten, wurden dabei immer komplexer. Die fortschreitende Miniaturisierung führte zu höheren Schaltgeschwindigkeiten sowie einer Reduktion der Versorgungsspannungspegel von ehemals 5 V auf heute teilweise unter 1 V. Die Konsequenz dessen ist, dass die Bausteine zunehmend hochfrequente Signale emittieren und gleichzeitig anfälliger für Störungen von außen werden.
Deshalb hat die Internationale Electrotechnical Commission (IEC) zwei Normen zur Charakterisierung von ICs veröffentlicht. Die IEC 61967 für die Störaussendung und die IEC 62132 für die Störfestigkeit. Teil 4 der IEC 62132 (IEC 62132-4) spezifiziert dabei die Direct Power Injection (DPI) Methode, die es ermöglicht, die Störfestigkeit einzelner IC-Pins gegenüber leitungsgeführten Störungen zu bestimmen. Auf Basis dieser Normen hat der Zentralverband Elektrotechnik- und Elektronikindustrie e.V. (ZVEI) die "Generic IC EMC Test Specification" erstellt. Diese Spezifikation beschreibt standardisierte Testverfahren zur Charakterisierung des EMV-Verhaltens von ICs. Dabei werden verschiedene IC-Typen betrachtet und deren Pins nach ihren Funktionen klassifiziert. Zum Beispiel wird zwischen Spannungsversorgungspins, digitalen und analogen Pins unterschieden. Eine weitere Klassifizierung geschieht in lokale und globale Pins. Globale Pins sind dadurch gekennzeichnet, dass sie an Konnektoren der Leiterplatte angeschlossen sind und sie damit potenziell größeren Störungen ausgesetzt sind. Globale Pins können sowohl an Daten- als auch an Versorgungsspannungsleitungen angeschlossen sein. Lokale Pins hingegen sind ausschließlich auf der Leiterplatte verdrahtet [4].
Ursprünglich wurden DPI-Tests hauptsächlich von Automobilzulieferern für die Transceiver-Bausteine von Bussystemen wie CAN und LIN gefordert. Die Leitungen dieser Systeme sind im gesamten Fahrzeug verteilt und dadurch besonders anfällig für elektromagnetische Störungen. Mit der zunehmenden Elektrifizierung und Digitalisierung im Automobilbereich wurden die elektrischen Systeme immer komplexer, was zu einem deutlichen Anstieg der Kabelbaumlänge führte. In modernen Fahrzeugen beträgt die durchschnittliche Kabelbaumlänge bereits 8 km [2]. Diese Entwicklung unterstreicht die wachsende Bedeutung von IC-Level EMV-Tests wie der DPI-Methode. Entsprechend werden diese Tests in immer mehr Produktanforderungen vorgeschrieben. Dadurch werden Halbleiterhersteller verpflichtet, ihre Produkte nach diesen Standards zu prüfen.

Zielsetzung

Im Rahmen der Thesis wird die Direct Power Injection Methode nach der IEC 62132-4 anhand eines Schaltreglers praktisch angewendet und analysiert. Ein Schwerpunkt der Arbeit liegt auf dem Design einer Leiterplatte, mit der ein Schaltregler untersucht und getestet werden soll. Die Leiterplatte umfasst neben dem Schaltregler auch die für die Testdurchführung notwendigen Koppelnetzwerke und Filterschaltungen. Beim Design der Leiterplatte ist darauf zu achten, dass diese für Hochfrequenzsignale geeignet ist. Das heißt, Leiterbahnen müssen impedanzkontrolliert sein und parasitäre Induktivitäten und Kapazitäten müssen vermieden werden. Eine parasitäre Kapazität kann sich zum Beispiel zwischen den Wicklungen einer Spule und

der Massefläche unter dieser bilden. Hierdurch wird das Hochfrequenzsignal an dieser Stelle kurzgeschlossen.
Zusätzlich zur Hardwareentwicklung wird ein Messaufbau nach Norm realisiert. Dabei ist vor allem das Überwachen der Funktionen des ICs während des Tests von Interesse. Ziel ist es, den Test automatisch durchzuführen und zu untersuchen, wie sich das Verhalten des ICs mit unterschiedlichen Koppelnetzwerken und Arbeitspunkten verändert.

Messaufbau

In Abbildung 1 ist schematisch der Messaufbau für einen DPI-Test dargestellt. Das hochfrequente Signal wird von einem Frequenzgenerator erzeugt und über eine 50 Ohm Leitung zu einem Verstärker geführt. Danach kommt ein Richtkoppler, mit dem die Vorwärtsleistung und die reflektierte Leistung gemessen werden. Hierfür wird ein nicht signifikanter Teil des Signals ausgekoppelt und mit einem Leistungsmesser erfasst [5].
Anschließend wird das Signal über einen 50 Ohm Stecker auf die Leiterplatte geleitet und mit einer impedanzkontrollierten Leiterbahn zum Koppelnetzwerk geführt. In der Norm IEC 62132-4 wird ein Koppelkondensator mit 6.8 nF vorgeschrieben. Zur Limitierung des Stroms kann ein Serienwiderstand hinzugefügt werden. Nach dem Koppelnetzwerk wird das 50-Ohm-System verlassen, weshalb der Kondensator möglichst nah an dem zu untersuchenden IC-Pin platziert werden muss, um ungewollte Reflexionen zu vermeiden.
Während des Tests wird das Device Under Test (DUT) in einem definierten Arbeitsmodus betrieben. Das heißt, je nach Funktion des Pins wird dieser mit einer DC-Spannung oder einem Signal beaufschlagt. Zum Schutz der Messausrüstung vor den hochfrequenten Störungen wird ein Tiefpassfilter in der Signalleitung implementiert. Dies gilt sowohl für die Versorgung und Eingangssignale als auch für die Überwachung des ICs. Ein weiterer wichtiger Aspekt des Messaufbaus ist die Überwachung des ICs, um Fehlfunktionen während des Tests feststellen zu können. Ein Beispiel ist die Überwachung der Ausgangsspannung eines Schaltreglers: Weicht diese mehr als 5 % von ihrem Sollwert ab, wird dies als Fehlfunktion interpretiert.

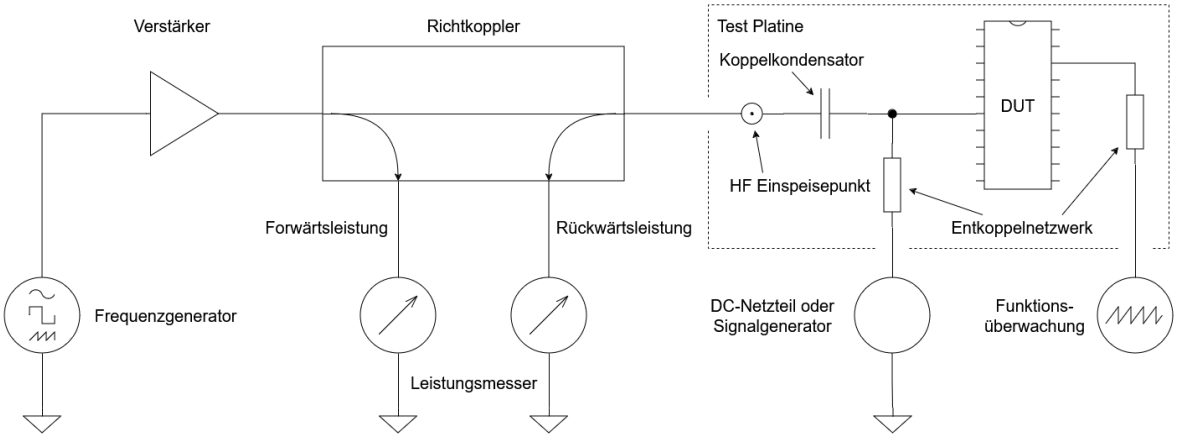


Abb. 1: Schematische Darstellung eines DPI-Messaufbaus [1]

Testdurchführung

Zu Beginn der Messung werden alle Geräte an die DUT-Leiterplatte angeschlossen und der IC auf korrekte Funktion überprüft. Danach wird systematisch das hochfrequente Störsignal eingespeist. Es wird mit der niedrigsten Frequenz von 150 kHz und einem geringen Leistungspegel, typischerweise 20 dB unterhalb der Zielleistung, begonnen. Wird ein globaler Pin getestet, beträgt die Zielleistung 37 dBm, bei einem lokalen Pin liegt sie zwischen 17 dBm und 23 dBm.
Die Leistung liegt immer mindestens für eine Sekunde an, während der die Funktion des ICs überprüft wird. Ist kein Fehler aufgetreten, wird die Leistung schrittweise um 0.5 dB erhöht, bis die Zielleistung erreicht ist oder bis ein Fehler auftritt [3]. Im Falle eines Fehlers wird der maximale Leistungspegel ohne Fehlfunktion erfasst und der Test bei der nächsthöheren Frequenz fortgesetzt. Die Sequenz wird so lange wiederholt, bis die maximale Frequenz von 1 GHz erreicht ist. Dies wird für alle zu testenden Pins wiederholt. In Abbildung 2 ist das Messergebnis eines Vin Pins eines Schaltreglers dargestellt. Bei niedrigen Frequenzen funktioniert der IC bei der maximalen Leistung, ab 520 MHz ist das erste Mal eine Fehlfunktion des ICs zu erkennen. Zwischen 700 MHz und 880 MHz musste die Leistung deutlich reduziert werden, damit der IC noch fehlerfrei funktioniert.

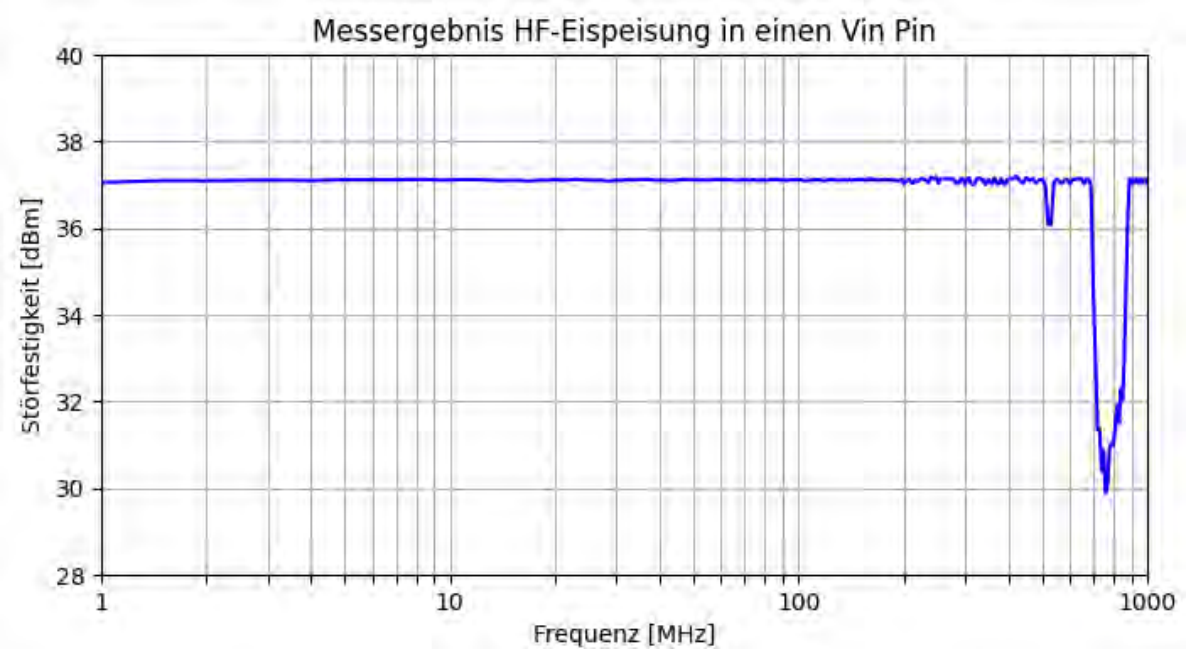


Abb. 2: Messergebnis HF-Einspeisung in einen Vin Pin [1]

Ausblick

Die systematische Charakterisierung der Störfestigkeit einzelner IC-Pins ermöglicht es, in frühen Entwicklungsphasen potenzielle EMV-Schwachstellen zu identifizieren und mit zusätzlichen Maßnahmen zu beheben. Für den Schutz sensibler Pins stehen dabei verschiedene Möglichkeiten zur Verfügung. So können durch die Integration zusätzlicher Filterkondensatoren hochfrequente Störsignale gedämpft werden. Dabei ist auf eine niederinduktive Anbindung der Kondensatoren zu achten, um deren Wirksamkeit auch bei hohen Frequenzen sicherzustellen. Eine weitere Möglichkeit ist die Verdrahtung kritischer Signalleitungen auf den Innenlagen der Leiterplatte, da die umgebenden Masselagen als Schirm funktionieren und somit die Einkopplung von Störsignalen reduzieren.
Für Halbleiterhersteller bietet die DPI-Methode den Vorteil, dass sie ihre Produkte standardisiert analysieren und Rückschlüsse auf das interne Design bezüglich Störfestigkeit ziehen können. Diese Erkenntnisse können in zukünftigen Neuentwicklungen berücksichtigt werden und zur Produktverbesserung beitragen.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Gordian Hense. Tesla spart bis zu 2 Kilometer Kabel und reicht neues Patent ein. https://www.carmart.ch/elektro/tesla-spart-bis-zu-2-kilometer-kabel-und-reicht-neues-patent-ein/, 12 2019.

[3] International Electrotechnical Commission IEC. *Integrated circuits - Measurement of electromagnetic immunity 150 kHz to 1 GHz - Part 4: Direct RF power injection method*. IEC, 2006.

[4] Wolfgang Pfaff et al. Generic IC EMC Test Specification Version 2.1. https://www.zvei.org/presse-medien/publikationen/leitfaden-generic-ic-emc-test-specification-version-21, 2017.

[5] Art Pini. HF-Richtkoppler – Grundlagen und effiziente Verwendung. https://www.digikey.de/de/articles/the-fundamentals-of-rf-directional-couplers-and-how-to-use-them-effectively?srsltid=AfmBOoovrtEsWjtNs1pjP1pEZVr0BnHVwpWdujIEcDKyurqWwrGM4HcN, 10 2019.

[6] M. Ramadani et al. The Electromagnetic Compatibility of Integrated Circuits—Past, Present, and Future. *IEEE Transactions on Electromagnetic Compatibility*, 51, 2009.

Optimierung von Regressions- und Lasttests in agilen Entwicklungsumgebungen

Erik Landgrebe **Dieter Morgenroth**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Mercedes-Benz Tech Innovation GmbH, Stuttgart

Diese Situation kennt jeder: Eine Handy-App geöffnet und die Mitteilung erhalten, dass man doch bitte auf die neuste Version updaten soll, um fortzufahren. Dies ist inzwischen Teil des Lebens und Alltag für alle, die sich im digitalen Raum bewegen. Doch wo kommen diese Updates her und wie wird sichergestellt, dass diese Updates auch funktionieren? Hier greift diese Arbeit ein und zeigt, wie Regressions- und Lasttests in agilen Entwicklungsumgebungen optimiert werden können.

Ein Update stellt eine neue Version derselben Anwendung dar, in welcher alte Probleme behoben oder neue Funktionen eingeführt werden. Diese neuen Versionen müssen noch während der Entwicklung getestet werden, bevor sie zum Endabnehmer ausgeliefert werden. Wenn die Entwicklungszeit immer schneller und in regelmäßigen Abständen neue Versionen erscheinen sollen, so muss das Testen auch in einem ähnlichen Zyklus stattfinden, wie es die Entwicklung macht. Wird eine neue Funktion beendet, so muss diese noch innerhalb desselben Zyklus getestet werden, so dass sie ohne Fehler funktioniert.

Updates können sich jedoch nicht nur auf Funktionen für Kunden, sondern auch die zu Grunde liegende Hardware beziehen. Diese muss auch, genauso wie die geänderten Dienste, dauerhaft getestet werden, sodass Fehler und möglich Schwachstellen möglichst schnell gefunden und ausgemerzt werden können. Nicht nur ein Update der Infrastruktur kann sich auf die Belastung der Hardware auswirken, sondern auch Veränderungen an der Anwendung können Funktionen im zugrundeliegenden Code ändern, sodass diese nicht offensichtliche Auswirkungen und folgen auf die Hardwareauslastung haben können. Ein Computer kann nur begrenzt viele Aktionen gleichzeitig ausführen. Was passiert, wenn dieser ans Limit kommt? Dies muss getestet und erprobt werden. [4]

Untertitel/Headline Stand der Forschung

Regressions- und Lasttests werden schon seit sehr vielen Jahren durchgeführt, wie zum Beispiel in der Automobilindustrie, wo es um Leben und Tod gehen kann, wenn die Software nicht so funktioniert, wie sie es soll. [2]

Als einen Weg der Eichung und Qualitätssicherung gibt es seit 2002 das International Software Testing Qualifications Board, kurz ISTQB, das verschiedene Niveaus an Testern ausbildet und weiterbildet, womit ein internationaler Standard gesichert werden soll. Sie stellen Test, die an ausgewählten Orten durchführt werden können, als auch generelle Fortbildungen. [3]

Die Wichtigkeit von Softwaretests nahm durch agile Arbeitsweißen deutlich zu. Auch kam ein Trend zu immer mehr Softwaretests, diese sollen aber professioneller werden, damit die Qualität von Software steigt. [1] Softwarequalität verbessert sich zum Beispiel, wenn man sich an den Deming-Zyklus, wie er in Abbildung 1 [1] gezeigt wird, hält, welcher einen von vielen Ansätzen widerspiegelt, mit dem man schon in der Planung von Tests die Qualität verbessern kann. Wenn die Tests besser werden, wird auch die Software, welche getestet wird, besser.

Abb. 1: Deming-Zyklus am Beispiel des Tests verdeutlicht [1]

Das Anpassen von Testen an die Anforderungen eines agilen Projektes wurden in den letzten Jahren weniger erforscht, seitdem agile Arbeitsweißen der Standard wurden. Das Testen musste sich hierbei aber stark anpassen, da kontinuierliche Integration und kontinuierlicher Auslieferung inzwischen in sehr vielen Projekten der Alltag sind. So wurde im Jahr 2020 im Bereich Testen im Agilen Umfeld der größte Weiterbildungsbedarf gesehen, wo die Bachelorarbeit anschließen soll. [5]

Untertitel/Headline Methode

Um Regressions- und Lasttests zu optimieren, werden in einem Projekt für ein großes, süddeutsches Industrieunternehmen schon bestehende Tests genommen und angepasst, oder auch neue Tests erstellt, die verschiedene Funktionen eines bereits weitgehend bestehenden Produktes testen, welches sich aber noch in einem SCRUM-Umfeld entwickelt, weiterentwickelt und verbessert wird. Es wird untersucht, was der Test bewirken soll, was er wirklich bewirkt und wie er verbessert werden kann, damit das Ziel erfüllt wird. Auch wird untersucht, wie mit abgeschlossenen Tests umgegangen werden soll, wenn sie erfolgreich waren oder auch fehlgeschlagen sind.
Primär soll eine Trial-and-Error Methodik angewendet werden, um so Tests zu optimieren. Während der Arbeit wird immer wieder ein Bezug zu verschiedensten Normen und Regeln genommen, wie zum Beispiel jenen des ISTQB, um diese einzuschätzen, zu bewerten, als auch möglicherweise anzupassen. Neben dem Anpassen von Tests wird auch zwischen verschiedenen Test-Werkzeugen verglichen, um auch hier Vor- und Nachteile zu finden und erwägen.
Dies wird alles durch eine prototypische Entwicklung geschehen, so dass ein Test entwickelt wird, dieser wird eingesetzt und untersucht, was sich wie verändert hat. Hierfür werden verschiedene Metriken gesucht und verwendet, anhand welcher man Tests vergleichen kann. Die Tests werden zwar primär auf einem Dienst laufen, werden aber für allgemeine Schlüsse auch diesbezüglich betrachtet, damit Optimierungen möglichst allgemein gelten. Ein Ziel wird zum Beispiel sein, die Testabdeckung zu verbessern. Um einen Überblick der Tests zu erhalten und um Veränderungen und deren Auswirkungen testen zu können wird zum Beispiel in einer Tabelle festgehalten, was geändert wurde, wie es sich ausgewirkt hat, als auch, welche Schlüsse man daraus ziehen kann.

Literatur und Abbildungen

[1] Manfred Baumgartner, Martin Klonk, Christian Mastnak, and Richard Seidl. *Agile Testing: der agile Weg zur Qualität*. Hanser, 2024.

[2] Christian Berger. Accelerating Regression Testing for Scaled Self-Driving Cars with Lightweight Virtualization – A Case Study. In *2015 IEEE/ACM 1st International Workshop on Software Engineering for Smart Cyber-Physical Systems*. IEEE, 2015.

[3] Renzo Cerquozzi et al. *Certified Tester Foundation Level Syllabus v4.0*. International Software Testing Qualifications Board, 2024.

[4] Peter Liggesmeyer. *Software-Qualität: Testen, Analysieren und Verifizieren von Software*. Spektrum, 2002.

[5] Mario Winter et al. Softwaretest in Praxis und Forschung. https://www.softwaretest-umfrage.de/2020/index.html, 2020.

Lean Management Prinzipien für den Einsatz im IT Demand Management

Alexander Leppich **Thomas Rodach**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma TRUMPF SE + Co. KG, Ditzingen

Einleitung

In der heutigen Zeit unterliegen Unternehmen einem ständigen Wandel, der sowohl durch technologische Entwicklungen als auch durch sich ändernde Geschäftsanforderungen getrieben wird. Der IT Demand Prozess, welcher die Schnittstelle zwischen den Fachbereichen und der IT Abteilung darstellt, ist zu einem immer komplexer werdenden Gebilde geworden. Diese Weiterentwicklungen können im Laufe der Zeit zu zahlreichen ineffizienten Flaschenhälsen, dezentralen Abläufen und redundanten Prozessen führen. Der Ansatz Lean Management welcher Prinzipien beinhaltet, um Prozesse in Unternehmen nachhaltig zu optimieren hat sich bereits in anderen Sektoren der Industrie als eine Art Best Practice etabliert und findet heutzutage auch zunehmend Anwendung in der IT. [5]

Definition und Ursprung

Der Ansatz "Lean Management"hat seinen Ursprung in der Fertigungsindustrie, insbesondere im Toyota-Produktionssystem, welches in den 1950er Jahren in Japan entwickelt wurde. Es handelt sich hierbei um einen kundenzentrierten Optimierungsansatz, dessen Ziel es ist, Prozesse effizienter zu gestalten, Verschwendung zu minimieren und den Kundenwert infolgedessen zu maximieren. Mit der Zeit wurde der Lean Ansatz über die Fertigung hinaus in verschiedenen anderen Branchen wie dem Gesundheitswesen, der Logistik und der IT angewendet, wo er ähnliche Effizienzsteigerungen und Prozessoptimierungen ermöglicht. [4]

Wertschöpfung und Verschwendung

Die wichtigste Metrik im Lean Management wird als „Wertschöpfung“ beschrieben, also alle Prozesse, Produkte und Ergebnisse, welche einen bestimmten Wert für die Organisation schaffen, von welchem der Kunde profitiert. [1] Alle restlichen Prozesse, welche keinen direkten Wert schaffen werden somit als verschwenderische Tätigkeiten dargestellt, da diese dem Kunden keinen Nutzen erbringen. Verschwendungen in einem Prozess zeigen sich hierbei oft in Form von Überproduktion, Wartezeiten, unnötigen Transporten, überflüssiger/redundanter Arbeit, lagernden Beständen, Bewegungen und Zeit, welche durch Fehlerbehebung genutzt werden muss. Diese Arten der Verschwendung lassen sich auch auf IT Prozesse übertragen, bei welchen Prozesse und Teilprozessschritte keinen direkten Wert für den Kunden generieren und somit als „verschwenderisch“ zu klassifizieren sind. [4]

Lean Management Prinzipien

Um nun einen Prozess im Rahmen des Lean Managements zu optimieren ist es zentral alle verschwenderischen Tätigkeiten entweder zu eliminieren oder in wertschöpfende Tätigkeiten für den Kunden umzuwandeln. [4] Hierbei stützt sich der Ansatz auf verschiedene Prinzipien, welche auf den gesamten Prozess als Top-Down und auf die einzelnen Tätigkeiten als Bottom-Up angewandt werden können. Zu den zentralen Prinzipien gehören: [4]

- Wertdefinition aus Kundensicht: Fokus auf wertschöpfende Aktivitäten, die einen klaren Nutzen für den Kunden schaffen.
- Wertstromanalyse: Detaillierte Analyse von Prozessen zur Identifikation und Eliminierung nicht-wertschöpfender Schritte.
- Fluss-Prinzip: Sicherstellung eines kontinuierlichen Arbeitsablaufs ohne jegliche Unterbrechungen.
- Pull-Prinzip: Steuerung der Produktion oder Prozesse basierend auf tatsächlicher Nachfrage des Kunden.
- Streben nach Perfektion (Kaizen): Kontinuierliche Verbesserung zur iterativen Optimierung von Prozessen und Anpassung kleinster Elemente zur Erreichung eines optimalen Zustandes.

Diese Prinzipien bilden die Grundlage um einen IT Demand Prozess, nach dem Lean Management Ansatz, schlank und kundenorientiert zu gestalten. [1]

IT Demand Prozess

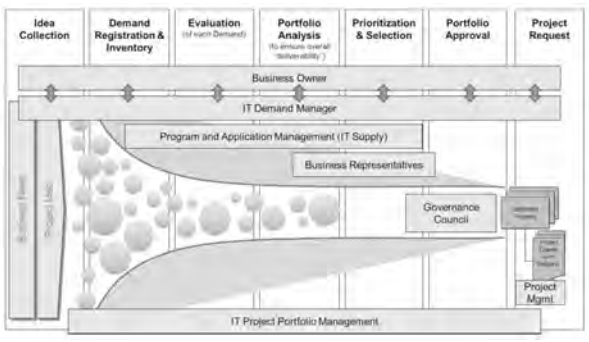


Abb. 1: Darstellung eines IT Demand Prozesses als IT-Projekttrichter [3]

Um die Anforderungen der Fachbereiche eines Unternehmens gezielt umzusetzen, wurde der IT Demand Prozess etabliert, welcher darauf abzielt, die Effizienz zu steigern und die Zusammenarbeit mit den IT Abteilungen zu optimieren. Dieser Prozess unterteilt sich oft, wie in Abbildung 1 dargestellt, in verschiedene Phasen von der Aufnahme der Ideen der Fachbereich bis hin zur Umsetzung durch die IT: [3]

1. Demandaufnahme: In dieser ersten Phase formulieren die Fachbereiche ihre zu erfüllenden Ideen mit ersten groben Details und einem vorläufigem Zeitplan.
2. Untersuchung des Demands: Das IT Demand Management prüft den eingegangenen Demand auf Vollständigkeit, bestimmt die Priorität sowie die Machbarkeit.
3. Technische und ökonomische Evaluation: Die IT Abteilungen untersuchen den Demand auf die tatsächliche technische Umsetzung, die Kosten und den Nutzen durch die Erstellung eines IT Projektes und wägen diese gegeneinander ab.
4. Genehmigung: Wurde ein Demand nun vollständig untersucht und ein vorläufiger Plan zur Allokation der Kapazitäten erstellt, folgt schließlich eine Genehmigung, durch welche ein Demand schließlich zu einem IT Projekt wird.

Entlang dieses Prozesses gibt es viele Schnittstellen, sowie Teilschritte, welche sich durch Fehler und Redundanzen negativ auf die Durchlaufzeiten der Demands auswirken können. Daher ist es notwendig Optimierungsansätze in den IT Demand Prozess zu integrieren um eine hohe Effizienz und Qualität zu gewährleisten. [3]

Methodik

Die Erforschung neuer Optimierungsansätze stützt sich auf eine umfassende Literaturrecherche sowie die Modellierung und Analyse verschiedenster Abläufe im IT Demand Prozess. Diese Auswertungen werden zusammen mit verschiedenen Experten des Prozesses evaluiert und auf Basis dessen werden schließlich Handlungsempfehlungen für die Implementation entwickelt.

Anwendungsbeispiel: Wertstromanalyse eines ITD Prozesses

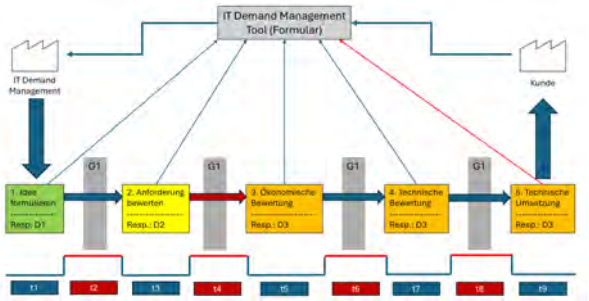


Abb. 2: Schematische Wertstromanalyse des ITD Prozesses [2]

Eine Anwendung der Lean Management Prinzipien auf den IT Demand Prozess kann, wie in Abbildung 2 vereinfacht dargestellt, durch eine Wertstromanalyse visualisiert werden. Wertschöpfende Tätigkeiten stehen hierbei grafisch im direkten Zusammenhang mit dem Nutzen für den Kunden und werden durch Pfeile entlang der Abbildung dargestellt. Alle Verschwendungen und Schwachstellen innerhalb des Prozesses werden hierbei ebenfalls als auffällige Elemente aufgezeigt, um sich auf diese in den nachfolgenden Optimierungsansätzen zu konzentrieren. [1]

Ansätze zur Optimierung des IT Demand Prozesses

Für die Erstellung von Optimierungsansätzen für den IT Demand Prozess, werden im Laufe der Bachelorarbeit die aufgenommenen Verschwendungen untersucht, um passende Maßnahmen aus der Theorie anzuwenden. Diese umfassen:

- Reduzierung von Mura (Unausgeglichenheit): Durch die Einführung eines gezielten Priorisierungssystems, welches auf wertschöpfenden Kriterien basiert, können Unausgeglichenheiten entlang des Prozesses vermindert werden indem

höher priorisierte Anforderungen zuerst bearbeitet und für einen kontinuierlichen Fluss an die nächste Einheit weitergegeben werden können. Wird diese Priorisierung transparent in den Prozess eingeführt, können Flaschenhälse durch vorausschauendes Verhalten minimiert werden und die Anzahl der bearbeiteten Demands erhöht werden. [4]

- Anpassung der Anforderungsformulare durch Poka-Yoke: Um Fehler in den einzelnen Prozessschritten, wie sie beispielsweise durch Missverständnisse entstehen können, zu vermeiden, sollten bereits bei der Gestaltung der jeweiligen Formulare Kontrollmechanismen integriert werden. Durch diese Integration von Poka-Yoke können auftretende Fehler unmittelbar erkannt und direkt korrigiert werden. [4]
- Einführung von Kaizen (Kontinuierlicher Verbesserung): Für eine nachhaltige Nutzung von Lean Management Prinzipien ist es notwendig, Verschwendungen bereits in kleinen Aufkommen aufzuzeichnen und diese dann zu beseitigen. Hierfür können regelmäßige Besprechungen mit allen Stakeholdern, welche sich nur auf die Optimierung konzentrieren dabei helfen, den Prozess schlank zu halten. [1]

Ausblick

Diese Methoden zur Prozessoptimierung basieren auf verschiedenen Theorien des Lean Managements. Im nächsten Schritt werden diese Ansätze nun im Austausch mit Experten validiert und infolgedessen in den aktuellen Prozess zur Verbesserung der Wertschöpfung und Verminderung von Verschwendungen integriert.

Literatur und Abbildungen

[1] Frank Bertagnolli. *Lean Management: Einführung und Vertiefung in die japanische Management-Philosophie*. Springer Gabler, 2018.

[2] Eigene Darstellung.

[3] S. Helmke and M. Uebel. *Managementorientiertes IT-Controlling und IT-Governance*. Springer Gabler, 2016.

[4] Marc Helmold. *Lean Management und Kaizen: Grundlagen Aus Fällen und Beispielen in Operations und Supply Chain Management*. Springer Gabler, 2023.

[5] A. Müller, H. Schröder, and L. von Thienen. *Lean IT-Management: Was die IT aus Produktionssystemen lernen kann*. Springer Science & Business Media, 2011.

Automotive Signal Sound Classification Using Modern Deep Learning Techniques

Kevin Phuc Hoang Luu **Thao Dang**

Department of Computer Science and Engineering, Esslingen University

Work carried out at Mercedes-Benz AG, Sindelfingen

Introduction

In the **Central System Integration (CSI)** department at Mercedes-Benz AG, the focus is on ensuring the reliability of telematics systems through stress, stability, and performance tests during continuous operation, all within the framework of test automation. The **CSI Testing** team conducts various tests to ensure that components meet the company's high quality standards. Examples of these components include power management, startup and shutdown behaviour, and audio and image processing. By performing stress tests, the team evaluates the stability, availability, and performance of the systems and works to reproduce any errors in a controlled lab environment.
Among these components, **signal sounds** are particularly significant in vehicles. These include not only functional indicators like blinker sounds but also critical warning sounds that help ensure driver safety and may even prevent accidents. This underscores the importance of reliable audio systems in modern vehicles.
The classification of these signal sounds falls under the domain of audio processing and is a topic that the CSI Testing team aims to incorporate into their framework.

Research Objectives

The primary goal of this research is to integrate automotive signal sound classification into an API that supports real-time processing. To achieve this, the main objectives are:

1. Identifying suitable models for the classification of signal sounds.
2. Developing a custom dataset by preprocessing and combining existing audio data.
3. Integrating the classification models as an API for real-time audio processing.

Audio classification is the process of analyzing and categorizing audio into predefined categories based on their acoustic features, including identifying various types of audio such as sound noise, musical notes, or other similar data [2].
The audio data can vary from environmental sounds to instruments to speech recognition. For detecting signal sounds, environmental sound classification is particularly relevant due to its similarity in audio patterns. Deep learning models commonly used for this task include Convolutional Neural Networks (CNNs) and transformers. This paper focuses on a CNN model, as the application of transformer models in this thesis is ongoing and not yet completed.

Creating the Dataset

To better understand the data, the CSI Testing team has already integrated an audio processing tool into their framework. This tool focuses on identifying the source of the entertainment device, such as Bluetooth, USB, or FM radio. The audio files are not typical content like music or navigation but are generated with distinct continuous frequency patterns. These patterns are necessary to decode the devices, as each device plays a specifically designed audio file to simulate actual playback. The tool tests if the connection from the entertainment device is working properly, hence the name **entertainment device sounds**.
Signal sounds, on the other hand, are real sounds played through the vehicle's speaker. While their original files are provided, they are unstructured, in contrast to entertainment device sounds. Examples of signal sounds include the blinker, seatbelt warning, or speed limit warning. These are the sounds that need to be detected and classified in the work of this thesis and are the actual task at hand.
Both audio processing tools serve different purposes. One detects entertainment device sounds, identifying the audio source, while the other focuses on detecting and classifying signal sounds. The key requirement

is that both programs must work simultaneously. Therefore, an important step is to create a custom dataset from the available audio samples.
To ensure consistency, all audio files are resampled, mixed down to a single mono channel, and normalized. Both signal sounds and entertainment device sounds undergo their own augmentation processes. For signal sounds, variation is introduced by decreasing the volume by specific amounts. Since entertainment device sounds are generated and already in perfect form, white noise is added to simulate static noise from device connections, ensuring that it does not affect the model. After augmentation, each signal sound is combined with every entertainment device sound, creating a Cartesian product of the two audio files. This process, along with the original files, creates a large dataset ready for model processing.
From the raw audio, a log-scaled mel-spectrogram is created, capturing the frequency spectrum of the audio signal over time. Delta features, which represent the rate of change in the spectrogram over time, are then extracted. This is crucial because the delta features help remove the continuous frequencies present in the entertainment device sounds, allowing the model to focus on the signal-specific features. Only the delta features are used as inputs to the CNN model, as the model does not learn directly from mel-spectrograms which is expected. Figure 1 shows the mel-spectrogram and the delta features of a blinker sound combined with a Bluetooth sound.

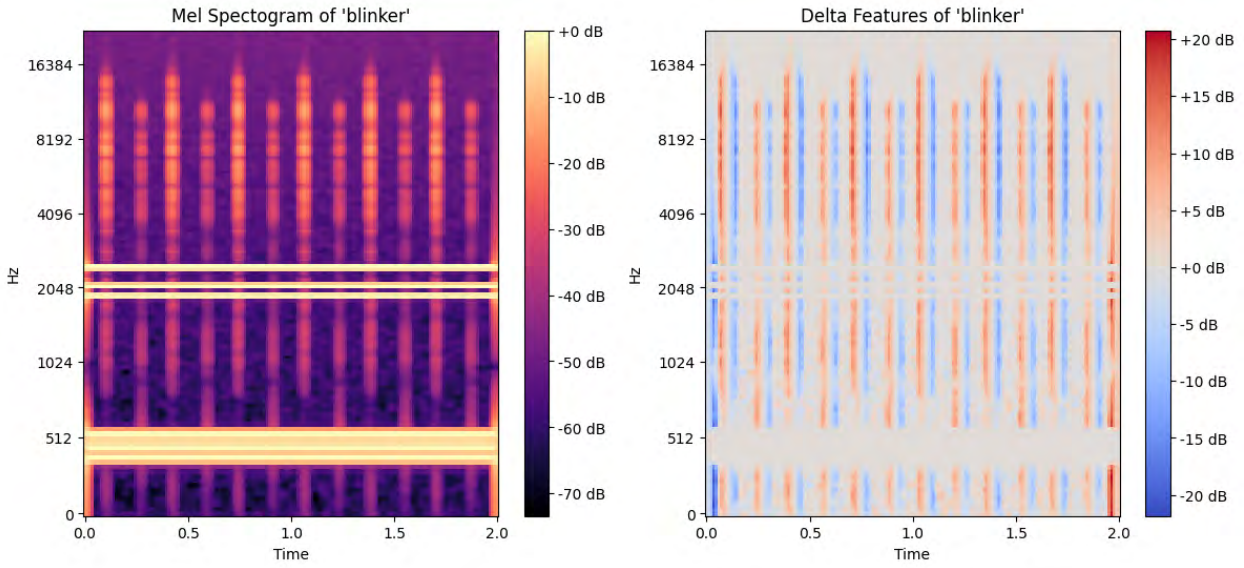


Fig. 1: Mel-Spectrogram and Delta Features of a Blinker Sound combined with a Bluetooth Sound [3]

Model

Convolutional Neural Networks are widely used in audio-related applications such as classification, speech recognition and music recommendation. In audio classification, CNNs have shown significant advancements, particularly in speech, music, and environmental sounds. Instead of using raw one-dimensional audio signals, CNNs often work with two-dimensional representations like spectrograms [4].
A typical CNN consists of an input layer, convolutional and pooling layers, fully connected hidden layers, and an output layer [1].
After creating the dataset, which contains the delta features as inputs for the CNN, the next step is to apply the CNN architecture for automotive signal sound classification. This architecture follows the same structure shown in Figure 2, with minor adjustments made to the layer sizes and filters.

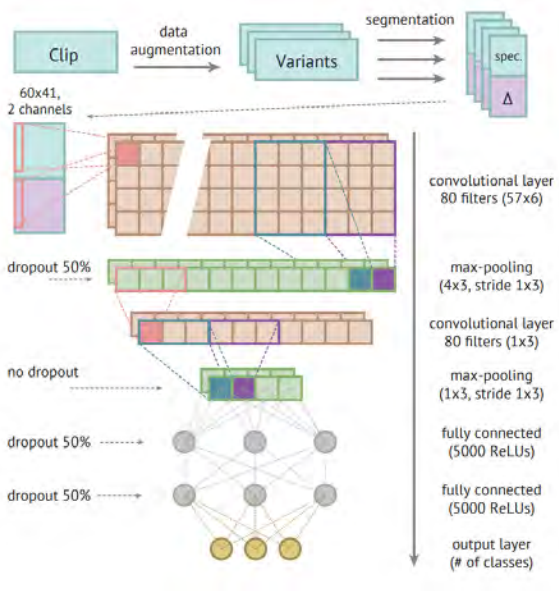


Fig. 2: CNN Model Architecture [1]

Evaluation

With a train and test accuracy of 99%, the model has effectively learned to classify the audio. While it could be argued that the model is overfitting, this is expected and even desirable given the limited data. Overfitting is acceptable in this lab environment because the conditions remain constant and the audio does not vary. The confusion matrix in Figure 3 for the CNN shows that most of the predicted labels are correct, with only a small margin of error. The most noticeable error is that the model predicted blinker nine times even though there was no sound. This is due to the distribution of the data after combining the audio data. Other than that, the model appears to classify the audio correctly nearly every time.

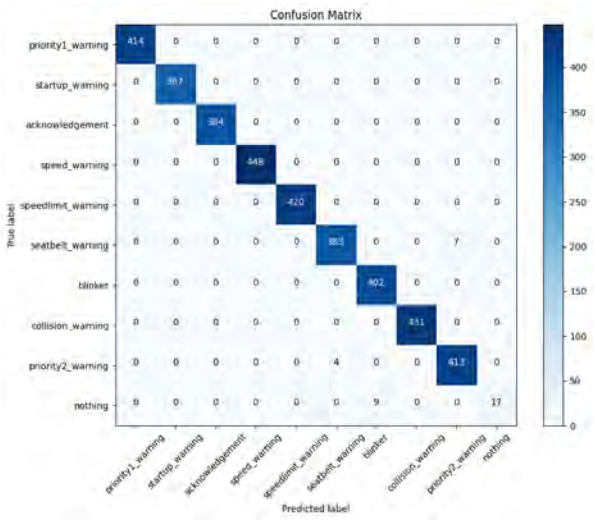


Fig. 3: Confusion Matrix Evaluation [3]

Outlook

Since CNNs are already performing effectively for the intended purpose, the next crucial step is to integrate the model into a real-time audio processing application. This application will process audio streams and return classification results along with confidence scores. Developing a real-time API dedicated to audio processing is a practical solution, as it can run on a central server, reducing complexity and facilitating integration into the framework used by the CSI Testing team. The API's independence from the framework's source code ensures flexibility and simplifies maintenance.

Following the integration, a potential next step would be to explore alternative models, such as transformers, and compare their performance to CNNs. This comparison could involve introducing new evaluation metrics to determine which model is more efficient and effective. Such an analysis would provide valuable insights into optimizing the classification process further.

References and figures

[1] Piczak Karol J. Environmental sound classification with convolutional neural networks. In *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2015.

[2] Shaoni Mukherjee. Audio Classification with Deep Learning | DigitalOcean. https://www.digitalocean.com/community/tutorials/audio-classification-with-deep-learning, 09 2024.

[3] Own representation.

[4] Khalid Zaman, Melike Sah, Cem Direkoglu, and Masash Unoki. A Survey of Audio Classification Using Deep Learning. *IEEE Access*, 11:106620–106649, 2023.

Konzeptionierung und prototypische Implementierung einer Statussynchronisation zwischen containerisierten Softwaremodulen innerhalb eines Betriebssystems

Moritz Malach **Mirko Sonntag**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Bosch Rexroth AG, Horb am Neckar

Einleitung und Problemstellung

Fortschritte in der Informationstechnologie verbinden die Automatisierung und Digitalisierung im Sinne von Industrie 4.0 und schaffen so vernetzte Produktionsanlagen bis hin zur Smart Factory. [2]. Grundstoffe für die Produktion werden in der Regel mit hydraulischen Großgeräten gefördert. Bosch Rexroth will mit seiner Produktlinie BODAS, Smarte Sensoren, Steuergeräte und einer Telemetrie Einheit, der Rexroth Connectivity Unit, kurz RCU, auch diese Geräte smart machen. Dazu wird die RCU in so genannten Off-Highway-Fahrzeugen an den CAN-Bus des Fahrzeugs angehängt [4]. Einzelne Module, welche auf dieser RCU ausgeführt werden, führen kritische Änderungen durch. Diese Module greifen auf geteilte Ressourcen zu, und können sich gegenseitig blockieren. Diese Synchronisierung erfordert aktuell externen Eingriff. Das Ziel dieser Arbeit ist es, die einzelnen Funktionen der RCU zu Synchronisieren und einen Sperrmechanismus für kritische Änderungen zu implementieren.

Grundlagen

Die Linux-basierte RCU ermöglicht durch den Open-Source-Ansatz die Entwicklung eigener Software Komponenten für den Kunden. Bosch Rexroth verwendet dazu Softwarepakete im Snap Format. Snaps ermöglichen die Ausführung von Programmcode in einer gekapselten Umgebung, einem Container, ohne dabei plattformabhängig zu sein. Des Weiteren wird dadurch ein sogenanntes „Rollback“ von Updates ermöglicht. Da es verschiedene Varianten der RCU mit zum Teil unterschiedlichen Prozessorarchitekturen gibt, ist die Plattformunabhängigkeit wichtig. Over-the-Air-Services, kurz xOTA, wie:

- SOTA, Software Updates
- FOTA, Firmware Updates auch als Flashen von Steuergeräten bekannt
- POTA, Parameter Lesen und Schreiben
- DOTA, Diagnose Toolkit

gehören zu den wichtigsten Funktionen der RCU. Diese haben erhebliche Auswirkungen auf gemeinsam genutzte Ressourcen wie beispielsweise das CAN-Interface der RCU [3].

Zielsetzung

In dieser Bachelorarbeit soll ein Konzept erarbeitet und eine prototypische Implementierung einer Synchronisation mit Hilfe eines Betriebsartenkoordinators durchgeführt werden. Dieser Betriebsartenkoordinator synchronisiert die Durchführung kritischer Änderungen der einzelnen Software-Module durch Freigaben. Da ein Update auch einen Neustart der RCU auslösen kann, soll der Sperrzustand über mehrere Boot-Vorgänger hinweg gesichert werden. Dieser soll sich in die bestehende Architektur integrieren und eine einfache Schnittstelle für anfragende Snaps zur Verfügung stellen. Da die RCU zum Teil sehr limitierte Ressourcen hat, soll der Betriebsartenkoordinator sparsam mit diesen Ressourcen umgehen.

Client und Server Ansatz

Durch mehrere Design-Iterationen ist in 1 ein Zustandsautomat für die anfragenden Snaps entstanden. 2 zeigt die eigentliche Logik des Betriebsartenkoordinators ebenfalls in einem Zustandsautomat.

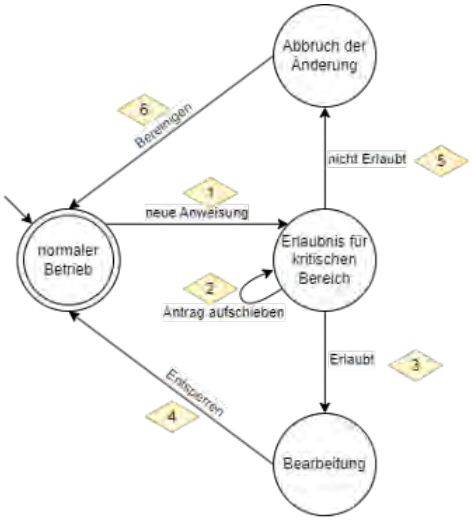


Abb. 1: Snap Ansicht [1]

Auf der Client-Seite, 1, startet das Modul im „normalen Betrieb“. Durch ein von außen ausgelöstes Ereignis erhält der Snap eine neue Anweisung (1). Nun fragt der Client den Server an, ob eine Änderung durchgeführt werden darf. Es gibt 3 mögliche Antworten, die der Client erhalten kann. Antrag aufschieben (2), eine Ausführung ist momentan nicht möglich und es kann später erneut versucht werden. Die Erlaubnis die Änderung durchzuführen (3), welche von der Meldung „Entsperren“ (4) gefolgt wird. Schließlich gibt es noch die Antwort „nicht Erlaubt“ (5). Diese wird gegeben, wenn der Anfragende Snap keine Berechtigung hat. Darauf folgt ein Bereinigungsjob (6).

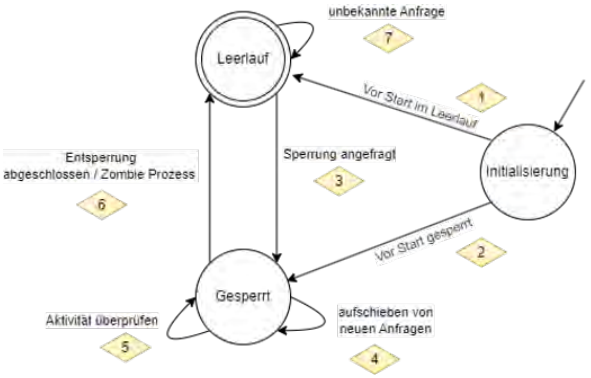


Abb. 2: Betriebsartenkoordinator Ansicht [1]

Der Server, 2, startet in der „Initialisierung“. Hier wird geprüft, ob vor dem Start des Systems bereits „gesperrt“ war. Je nach gespeichertem Zustand wird in den Sperrzustand übergegangen (2), oder in den Leerlauf (1). Wird nun eine Sperrung angefordert (3), wird das System gesperrt. Werden hier neue Anfragen erhalten, wird ein „Antrag aufschieben“ (4) versendet. Zu dem erfolgt eine Aktivitätsüberprüfung (5) des sperrenden Snap Moduls. Wenn keine Aktivität festgestellt wird, oder wenn der sperrende Snap ein „Entsperren“ sendet, wird die Sperrung aufgehoben (6).

Fazit und Ausblick

Das in der Bachelorarbeit konzipierte und prototypisch implementierte Modul ermöglicht die serielle Bearbeitung von Anfragen. Die für das Modul entwickelten Tests zeigen eine Ausführung einer Änderung, nur unter Bedingungen, die für das System als sicher gelten. Eine zukünftige Version könnte eine parallele Abarbeitung ermöglichen, wenn die zwei Module nicht auf gemeinsame Ressourcen zugreifen. Ebenso könnte eine Priorisierung eingeführt werden, die es einem bestimmten Snap erlaubt, die Sperrung exklusiv zu erhalten, wenn mehrere Anfragen gestellt werden.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Deutsches Institut für Normung. Was ist Industrie 4.0? https://www.din.de/de/forschung-und-innovation/themen/industrie4-0/was-ist-industrie-4-0, 2024.

[3] Bosch Rexroth. BODAS Connectivity Unit RCU. https://www.boschrexroth.com/de/de/media-details/4594dde8-66e8-4fa9-859e-7e11b2ef034e, 04 2022.

[4] Bosch Rexroth. BODAS HARDWARE. https://www.boschrexroth.com/de/de/transforming-mobile-machines/elektronifizierung-und-iot/bodas-hardware/, 2024.

Methodenentwicklung zur Überwachung und Analyse von Ergebnisdaten eines Fehlerabstellprozesses in einer heterogenen IT Landschaft

Tim Mencin **Clemens Klöck**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Mercedes-Benz AG, Sindelfingen

Einleitung

Moderne Fahrzeuge integrieren komplexe elektronische Systeme, die essenziell für Sicherheit, Komfort und Funktionalität sind. Die Fahrzeugproduktion erfordert umfangreiche Tests, die große Datenmengen generieren. Diese Daten werden in einer heterogenen IT-Landschaft verarbeitet, wobei Verzögerungen oder Fehler den Fehlerabstellprozess und die Qualitätssicherung beeinträchtigen können. Eine effiziente Überwachung der Datenflüsse ist daher wichtig für die Produktionsqualität.

Zielsetzung der Arbeit

Das Ziel dieser Arbeit ist die Entwicklung einer Methode, mit der Ergebnisdaten applikationsübergreifend verfolgt und in einem Dashboard visualisiert werden können. Zusätzlich werden die Daten analysiert, um mithilfe definierter Grenzwerte Engpässe zu erkennen. Auf dieser Basis werden Handlungsempfehlungen für Mercedes-Benz erarbeitet.

Vorgehen

Der Fehlerabstellprozess wurde zunächst analysiert, um die beteiligten Systeme, Schnittstellen und Datenflüsse zu identifizieren. Dabei lässt sich der Datenfluss in folgende Komponenten unterteilen, zu sehen in Abbildung 1:

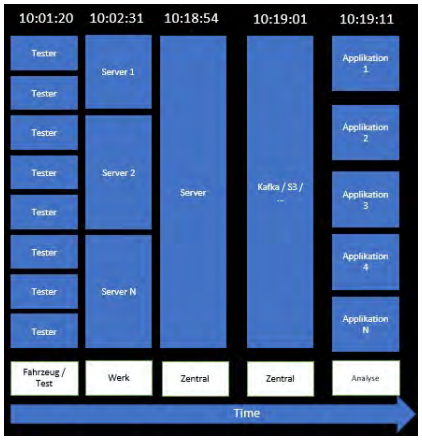


Abb. 1: Datenfluss des Fehlerabstellprozesses [2]

Zunächst werden die Tests der Fahrzeugelektronik für jedes Fahrzeug einzeln durchgeführt. Die dabei generierten Ergebnisdaten werden werksspezifisch gesammelt und anschließend zentral für alle Werke in der AWS Cloud zusammengeführt und gespeichert. Kafka, eine Daten-Streaming-Technologie, benachrichtigt die Konsumenten, wie etwa Analyseanwendungen, dass die Daten zum Abruf bereitstehen. Im letzten Schritt werden die Ergebnisdaten von den Analyseanwendungen entweder On-Demand oder eventbasiert abgerufen. Kafka stellt hierfür eine URL bereit, die einen direkten Download aus der AWS Cloud ermöglicht, wo die Daten in einer S3-Datenbank gespeichert sind.

Überwachungsstrategie

Für die Überwachungsmethode ergeben sich fünf Ansätze:

- Netzwerküberwachung
- Distributed Tracing
- Logging

- Eigene Lösung

Netzwerküberwachung

Die Netzwerküberwachung ist eine Methode zur Analyse des Datenverkehrs in Hochgeschwindigkeitsnetzwerken. Besonders bewährt hat sich die sogenannte Flow-Überwachung, die auf der Aggregation und Analyse von Datenflüssen anstelle einzelner Pakete basiert. Diese Technik nutzt Protokolle wie NetFlow oder IPFIX, um die Netzwerkdaten effizient zu exportieren und zu analysieren. [3]

Logging

Logging ist eine Methode zur Protokollierung von Ereignissen in IT-Systemen. Es zeichnet detaillierte Informationen über einzelne Ereignisse wie Fehlermeldungen, Benutzeraktivitäten oder Systemzustände auf. Dies dient der Fehlersuche, Analyse und Überwachung. [1]

Distributed Tracing

Distributed Tracing ermöglicht die Nachverfolgung von Abläufen in komplexen, verteilten Systemen wie Mikroservice-Architekturen. Es verknüpft Ereignisse über verschiedene Dienste hinweg und stellt so Kausalitätsbeziehungen her. Diese Methode wird verwendet, um Probleme wie hohe Latenzen zu erkennen und deren Ursachen zu analysieren. [1] Nach Evaluierung erwies sich Distributed Tracing als die passendste Lösung für den Anwendungsfall, da es eine umfassende Nachverfolgung der Ergebnisdaten ermöglicht – vom Teststart über die Verarbeitung durch verschiedene Microservices bis hin zur Analyseanwendung über mehrere Systeme hinweg. Allerdings konnte diese Methode im Rahmen der Arbeit nicht umgesetzt werden und wird daher lediglich als Handlungsempfehlung für Mercedes-Benz vorgeschlagen.

Eigene Lösung

Im Rahmen dieser Arbeit wurde eine eigenständige Lösung zur Überwachung des Datenflusses implementiert. Anstelle eines Agenten, der Trace-Daten der Systeme auswertet, wurde eine alternative Methode entwickelt, um die Nachverfolgung der Ergebnisdaten über verschiedene Microservices und Systeme hinweg zu ermöglichen. Dies war erforderlich, da weder Trace-Daten von den betroffenen Systemen erzeugt werden noch die Installation eines Agenten auf allen Systemen möglich ist. Die Nachverfolgung basiert auf Zeitstempeln, die entweder bereits in den Ergebnisdaten enthalten sind oder als Metadaten hinzugefügt werden. Die Zeitstempel werden durch zwei Hauptquellen bereitgestellt: den Streamingdienst Kafka und das Analysetool bzw. die Konsumenten. Die von Kafka bereitgestellten Zeitstempel decken den Zeitraum vom Teststart am Fahrzeug bis zur Benachrichtigung in Kafka ab. Die Zeitstempel des Analysetools umfassen den Zeitraum von der Benachrichtigung durch Kafka über neue Daten bis zum Empfang der Ergebnisdaten durch das Analysetool. Dieser Datenfluss wird in einem Sequenzdiagramm (siehe Abbildung 2) veranschaulicht, das die einzelnen Schritte und Zeitpunkte detailliert darstellt.

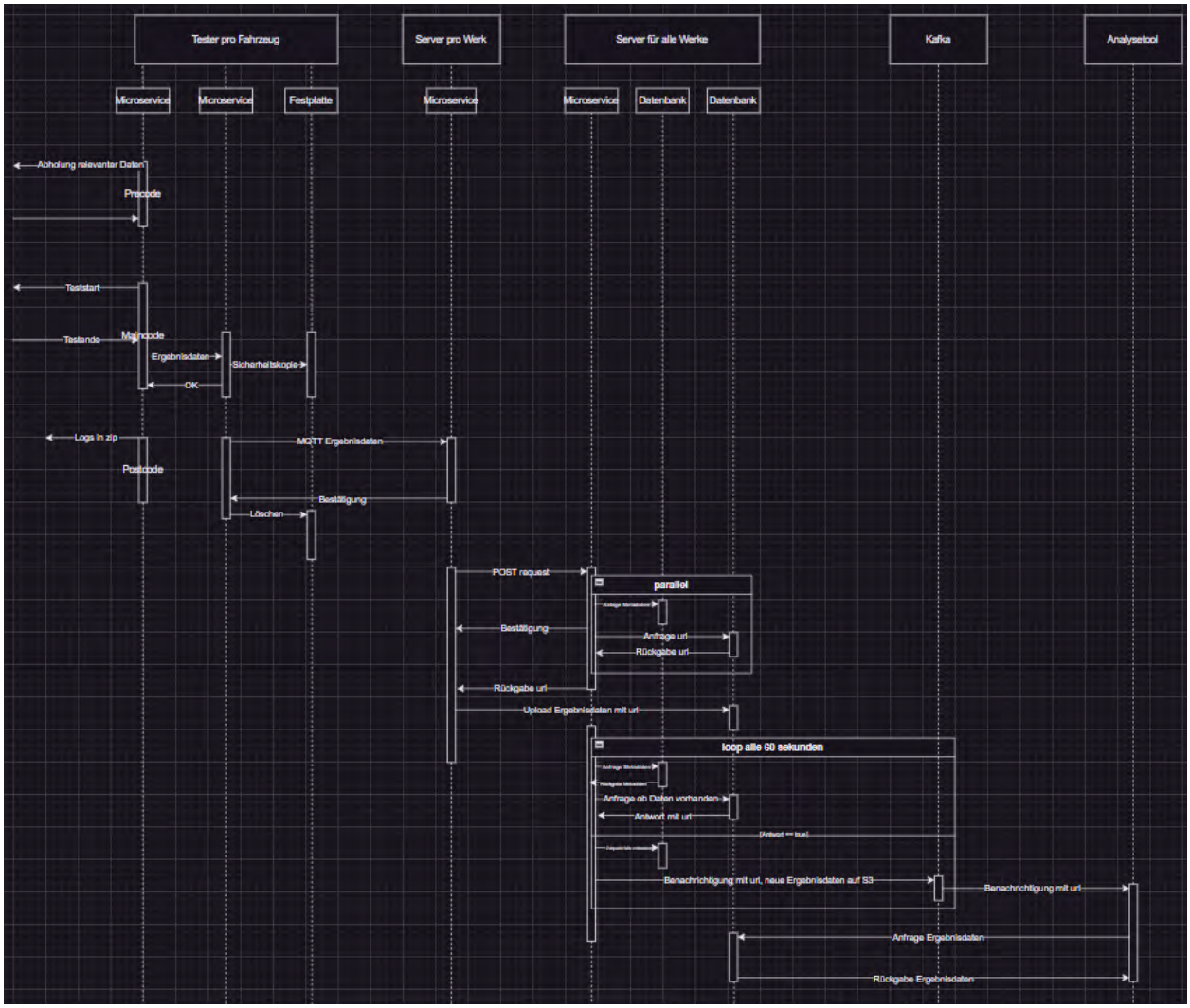


Abb. 2: Sequenzdiagramm des Fehlerabstellprozesses [2]

Identifikation relevanter Zeitstempel

Im nächsten Schritt wird analysiert, welche Zeitstempel benötigt werden, um den gesamten Prozess lückenlos abzubilden. Die Zeitstempel werden dabei in zwei Kategorien eingeteilt:

- Wünschenswert: Zeitstempel, die eine lückenlose Überwachung des Prozesses ermöglichen.
- Technisch realisierbar: Zeitstempel, die mit den vorhandenen Systemen und Schnittstellen tatsächlich erfasst werden können.

Wie bereits erläutert, können die meisten benötigten Zeitstempel über Kafka abgerufen werden. Dies ermöglicht jedoch ausschließlich die nachträgliche Analyse des Datenflusses. Für den Fall, dass Ergebnisdaten verloren gehen oder in einem System festhängen, wäre es notwendig, die Zeitstempel direkt von den jeweiligen Systemen zu beziehen. Diese Anforderung kann im Rahmen dieser Arbeit jedoch nicht umgesetzt werden und würde den Einsatz von Distributed Tracing erfordern, wie zuvor beschrieben.

Softwarearchitektur

Vor der praktischen Umsetzung wurde zunächst die Softwarearchitektur entwickelt. In Zusammenarbeit mit Softwarearchitekten und unter Berücksichtigung der definierten Anforderungen entstand der folgende Entwurf (siehe Abbildung 3):

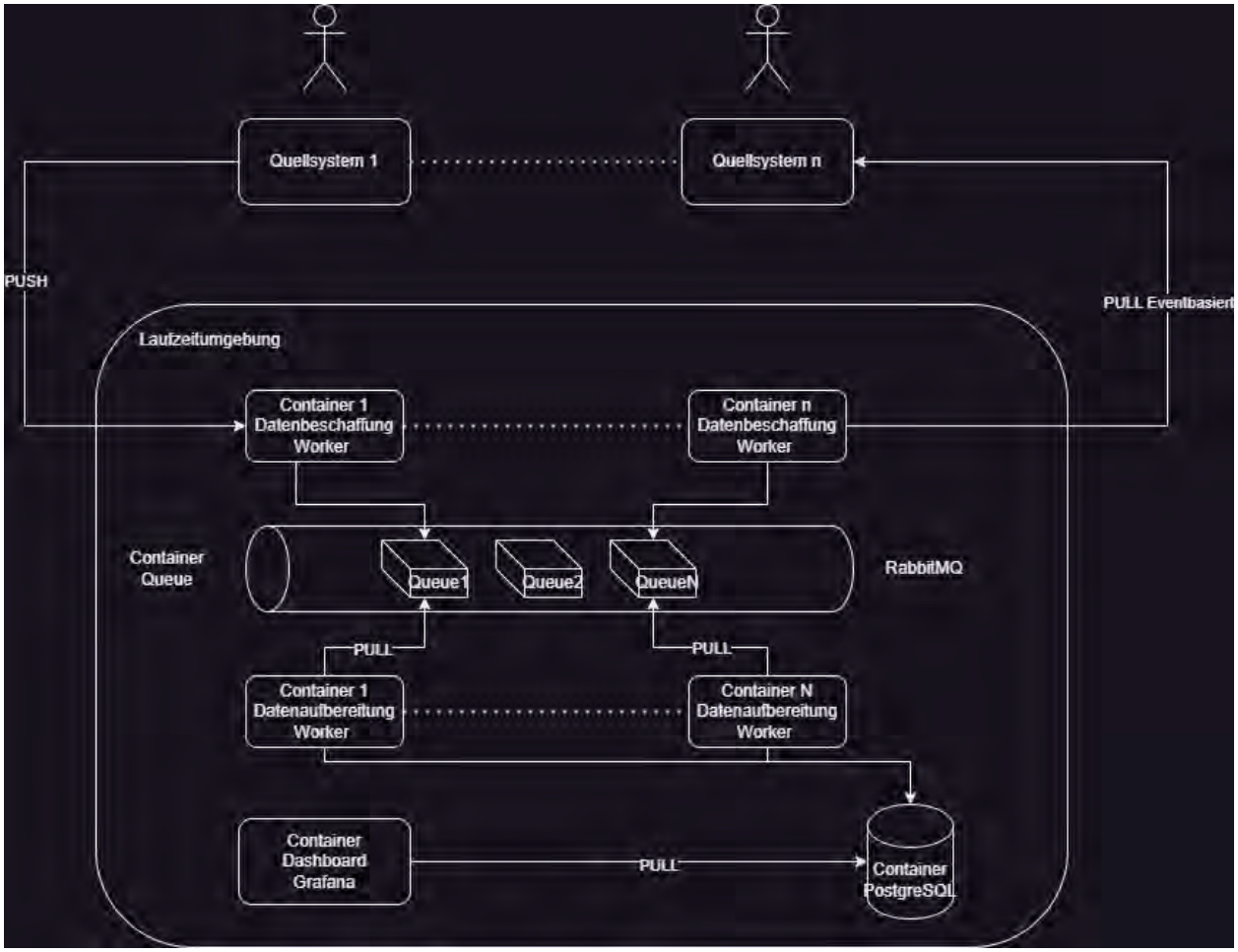


Abb. 3: Softwarearchitektur [2]

Der Entwurf gliedert sich in die folgenden Bereiche: Datenbeschaffung, Datenaufbereitung, Datenspeicherung und Datenvisualisierung. Diese Unterteilung lehnt an den Ansatz der Business Process Intelligence an, einer Methode zur „kontinuierlichen Analyse und anschließenden Optimierung bestehender Geschäftsprozesse." [5]. Der in dieser Masterarbeit beschriebene Prozess weist identische Anforderungen und Ziele auf, nämlich die vollständige Überwachung des Datenflusses in einer komplexen, heterogenen IT-Landschaft.

Umsetzung

Die Umsetzung der entworfenen Softwarearchitektur wurde zunächst lokal durchgeführt. Nach erfolgreicher Realisierung erfolgte im nächsten Schritt die Migration auf eine interne Plattform von Mercedes-Benz. Dabei wurde die zuvor entwickelte Docker-Umgebung in ein Kubernetes-Cluster integriert.

Datenbeschaffung

Im Rahmen der Datenbeschaffung wurde für jedes Quellsystem ein Python Docker Container implementiert. Dadurch werden Abhängigkeiten vermieden und eine leichte Skalierung ist gewährleistet, sollte es in Zukunft zu einer Erweiterung der Quellsysteme kommen. In dem Docker Container läuft ein in Python geschriebenes Programm, welches sich die Daten auf Event Basis per PULL-Prinzip [4] beschafft und anschließend auf eine Queue legt. Durch die Verwendung einer Queue wird die Datenverarbeitung von der Datenbeschaffung entkoppelt. Die Realisierung der Queue erfolgt mittels RabbitMQ. Auch hier wird für jedes Quellsystem eine eigene Queue verwendet. Zudem gibt es Quellsysteme, welche die Daten per PUSH-Prinzip [4] auf die Queue legen.

Datenaufbereitung

In der Folge werden die Daten von einem weiteren Python-Programm, dem sogenannten Worker, von der Queue abgeholt und aufbereitet. Hierbei kommt das

PULL-Prinzip zum Einsatz. Für jede Queue existiert ein entsprechender Worker. Sowohl die RabbitMQ als auch der Worker werden in einem Docker-Container ausgeführt. Die Aufbereitung umfasst die Extraktion der Zeitstempel aus den Daten sowie die Konvertierung der Zeitstempel in ein einheitliches Zeitformat gemäß dem ISO- und dem UTC-Standard. Zudem werden die Differenzen der Zeitstempel berechnet, um die Dauer der Ergebnisdaten von System oder Microservice zu nachfolgendem System/Microservice darzustellen.

Datenspeicherung

Die Zeitstempel werden in einer PostgreSQL Datenbank persistent abgespeichert. Diese läuft auch in einem Docker Container. PostgreSQL ist eine relationale Datenbank. Für jeden Meilenstein des Datenflusses gibt es eine Spalte, wo ein Zeitstempel abgespeichert wird. Zur eindeutigen Identifikation werden noch Vin (Vehicle Identification Number) und Werk hinterlegt.

Datenvisualisierung

Die Visualisierung der Zeitstempel erfolgt in einem Grafana-Dashboard. Grafana wird in DevOpsUmgebungen häufig eingesetzt, um Systeme in Echtzeit zu überwachen, Anomalien zu erkennen und historische Trends zu analysieren. Das Dashboard wird auf Basis der Anforderungen der Nutzer erstellt und ist so konzipiert, dass es durch Self-Service-Funktionen flexibel anpassbar ist, um den individuellen Bedürfnissen der Anwender gerecht zu werden.

Ausblick

Im weiteren Verlauf der Arbeit ist die Implementierung von Flask-Backends für die Python-Programme geplant, welche die Datenübertragung zwischen der Queue und den Systemen übernehmen. Ziel ist es, eine dauerhafte und stabile Laufzeit der Anwendungen sicherzustellen. Nach der Backend-Implementierung erfolgt die Auswertung der erfassten Daten, um potenzielle Engpässe im Datenfluss zu identifizieren. Dies ermöglicht eine gezielte Eingrenzung von Verzögerungsursachen auf spezifische Systeme, wodurch notwendige Verbesserungsschritte geplant werden können. Darüber hinaus soll der Einsatz von Künstlicher Intelligenz getestet und auf Umsetzbarkeit geprüft werden, beispielsweise für die Generierung von SQL-Abfragen mittels natürlicher Sprache. Nach Abschluss der Arbeit ist eine Erweiterung der Datenflussüberwachung auf zusätzliche Datentypen sowie die Integration neuer Quellsysteme und Konsumenten vorgesehen. Abschließend werden Handlungsempfehlungen an Mercedes-Benz formuliert.

Literatur und Abbildungen

[1] Andre Bento, Jaime Correia, Ricardo Filipe, Filipe Araujo, and Jorge Cardoso. Automated Analysis of Distributed Tracing: Challenges and Research Directions. *Journal of Grid Computing*, 2021.

[2] Eigene Darstellung.

[3] Rick Hofstede, Pavel Čeleda, Brian Trammel, Idilio Drago, Ramin Sadre, Anna Sperotto, and Aiko Pras. Flow Monitoring Explained: From Packet Capture to Data Analysis With NetFlow and IPFIX. *IEEE Commun. Surv. Tutorials 16 (4), S.2037-2064*, 2014.

[4] J. Martin-Flatin. Push vs. Pull in Web-Based Network Management. *IEEE*, 1999.

[5] Johannes Schobel. Business Process Intelligence Aktueller Stand und neue innovative Ansätze zur intelligenten Prozessanalyse, 2011.

Small Language Models in Intelligent Vehicle Assistants

Ilias Mirweis **Jürgen Koch**

Department of Computer Science and Engineering, Esslingen University

Work carried out at TomTom N.V., Amsterdam, Netherlands

Introduction

With the ever increasing digitization and integration of artificial intelligence (AI), intelligent vehicle assistants (IVAs) are gaining revived interest, because applications like ChatGPT are setting new expectations for that product category. These technologies allow or enable intuitive human-machine interaction. Small Language Models (SLMs) appeared as a key technology, which could offer robust language processing capabilities through their compact architecture and low resource requirements. Consequently, SLMs provide a foundation for implementing IVAs, which are more conversational and generally more capable in mobile devices, embedded systems, and other hardware contexts efficiently.

The central challenge in implementing systems like this lies in the selection and optimization of the underlying inference engines and environment and thus this results in upfront hardware purchasing considerations. Those factors form the technical backbone for executing language models and strongly influence performance, scalability, and resource utilization.

This bachelor's thesis includes a systematic analysis of controllable technical parameters and their impact on performance metrics in the context of using SLMs to build IVAs. The goal is to develop a benchmarking framework that assesses the efficiency, accuracy, and resource usage of these environment and metrics, providing a solid basis for practical integration decisions. This work is being carried out in collaboration with TomTom International B.V., a leading provider of navigation and mapping technologies. Beyond technical analysis, the study aims to provide insights into the scalability and optimization potential of Small Language Models in Intelligent Vehicle Assistants, contributing significantly to the development of resource-efficient and high-performance AI solutions.

Background and Motivation

Intelligent vehicle assistants are one of the most interesting advancements in AI, finding applications in a broad range of domains such as navigation, voice-controlled systems, and personalized services. Natural Language Processing (NLP) is an important component of these systems, allowing seamless and context-aware interactions. The focus of my thesis is on benchmarking SLMs to evaluate their performance across different hardware platforms and configurations. One approach to do so is using quantization techniques, which is a way to reduce the size and computational requirements of a machine learning model. Their impact on efficiency metrics such as latency, throughput, and memory consumption is measured. Quantization is particularly relevant for reducing the memory footprint and latency of neural network models, often achieving reductions between four to eight times in practice while maintaining acceptable accuracy [3]. In addition, quantization techniques are crucial for reducing computational costs and improving model deployment efficiency, especially on resource-limited devices [5]. With detailed benchmarking scripts, data on CPU and GPU performance was collected, providing a solid foundation for analyzing the trade-offs between model performance and resource usage in the context of intelligent vehicle assistants.

Inference engines are frameworks that execute machine learning models, which allows them to make predictions or inferences based on input data [1]. They handle tasks such as managing computations, allocating resources, and optimizing performance for real-time execution. These engines can run on a variety of devices, from powerful GPUs to embedded systems, and they are a critical component in deploying machine learning models effectively. They play a pivotal role in the real-time execution of these models. There are two types of inference engines used in this thesis: The first one is the Hugging Face Transformers library, known for its flexibility and precision, and the other is ONNX, which provides optimized performance for devices like mobile phones and embedded systems. The key challenge is finding the right balance between computational power, memory usage, and accuracy. This thesis aims to evaluate and identify which of these engines is better suited for Intelligent Vehicle

Assistants (IVAs) in different hardware setups and use cases.
The importance of this research arises from the growing demand for language models that must be both powerful and resource efficient. This systematic comparison of engines will yield valuable insights into which technologies are best suited for specific requirements and how existing systems can be made more efficient.

Objectives

The goal of this thesis is to develop a benchmarking framework that deeply evaluates the performance of different parameters. The comparison focuses on latency, accuracy, and resource consumption. The core research questions include:

1. How do different quantization techniques, such as dynamic quantization, FP16, and INT8, affect performance?
2. What differences arise when utilizing CPU and GPU hardware?
3. Which inference engine offers the best combination of scalability and efficiency for various application scenarios?

The benchmarking framework will measure not only the technical performance metrics of the engines but also their practical adequacy for IVA deployments. The goal is to provide a firm basis for integrating SLMs into real-world systems while identifying optimization opportunities.

Methodology

The evaluation will proceed through several steps. First, a consistent data set for comparison will be established using standardized runs. These runs represent typical IVA tasks, such as conversations, commands, and analysis. Two different inference engines will then be configured with identical SLMs. Quantized variants of the models will be used to analyze the effects of different precision levels.
Some attention might be given to trying various quantization techniques, as seen in Figure 1.For instance, one such technique is Quantization-Aware Training (QAT), which would enhance model efficiency without significant accuracy loss. The process of QAT can be visualized in the diagram below, which illustrates the steps involved in taking a pre-trained model, applying quantization, and then retraining or finetuning it with training data to obtain a quantized model [2]:

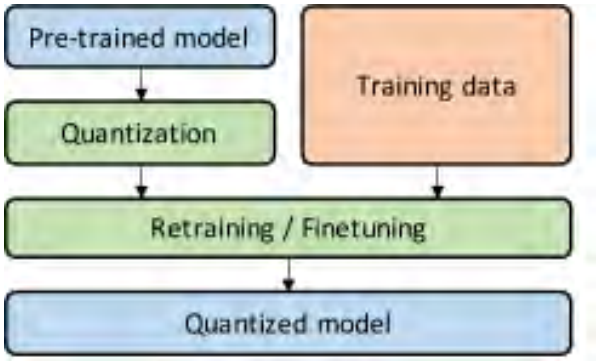


Fig. 1: Diagram showing the process of quantization-aware training, including the steps of applying quantization to a pre-trained model, retraining with training data, and producing the quantized model. [2]

As shown, the process begins with a pre-trained model, which is a model that has already been taught how to do a certain task. Then, we apply quantization, which is a way to make the model smaller and faster by reducing the amount of detail in its calculations. After that, we retrain or fine-tune the model so it can still perform well even with the reduced detail. This step is important to make sure that the model's accuracy isn't affected too much. By doing this, the model becomes more efficient without losing much of its ability to do the task well, unlike simpler methods that just make the model smaller without any extra training.

Results and Discussion

Initial results indicate that quantization techniques significantly impact engine performance. While some quantization methods on GPUs facilitates substantial performance improvements, some other quantization methods offers significant memory savings.
The transformer-based inference engine, which is from the HuggingFace library, demonstrates versatility but exhibits a strong dependency on high-performance hardware. In contrast, the ONNX Runtime delivers comparable accuracy with significantly lower resource consumption, making it particularly suitable for mobile and embedded systems. These differences underscore the importance of application-specific optimization of inference engines.
Furthermore, this setups allows me to prove or disprove hypothesis. With the amounts of metrics, it is possible to test certain use cases and scenarios, which is needed due to the variable nature of the project. It is immensely important to test to ensure comprehensive validation.

Fig. 2: Scatter plot illustrating the relationship between the number of input tokens and the time to the first token, indicating that there is no significant correlation between these two variables. [4]

Figure 2 illustrates the relationship between the number of input tokens and the time to the first token, demonstrating the hypothesis that there is no clear correlation between these variables.
As seen in Figure 2, one hypothesis was that there is no correlation between number of input tokens and time to first token. With the amount of metrics, this can be plotted to check whether statistical proofing is necessary or a sanity check is enough.

Outlook

The findings provide a solid foundation for further development of Language Models. Hybrid approaches that use and combine the strengths of both engines could provide even greater efficiency in the future. Furthermore, exploring emerging hardware platforms like Neural Processing Units, which is special hardware for machine learning use cases, hold great promise for further research. In the long term, the benchmarking framework can contribute to establishing efficient and scalable AI systems that meet the growing demands of modern edge-AI applications.

References and figures

[1] Jay E. Aronson. *Expert Systems*. Encyclopedia of Information Systems, 2003.

[2] Intel Corporation. Quantization in Intel Extension for Transformers. https://intel.github.io/intel-extension-for-transformers/latest/docs/quantization.html, 2022.

[3] Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W. Mahoney, and Kurt Keutzer. A Survey of Quantization Methods for Efficient Neural Network Inference. https://arxiv.org/abs/2103.13630, 06 2001.

[4] Own representation.

[5] Lu Wei, Zhong Ma, Chaojie Yang, and Qin Yao. Advances in the Neural Network Quantization: A Comprehensive Review. *Applied Sciences*, 14, 2024.

Deep learning methods for estimating focal length from a single image

Georgios Mitrakas **Markus Enzweiler**

Department of Computer Science and Engineering, Esslingen University

Work carried out at Spleenlab GmbH, Jena

Introduction

Calibrated cameras are essential for machine vision applications such as 3D reconstruction, depth estimation, and object localization. The Checkerboard calibration is arguably the most well-known and widespread technique for camera calibration, but it is time consuming and requires access to the physical camera. Automatic calibration methods that are fast and do not require access to the physical camera or checkerboard patterns are the focus of this work.
The industry standard pinhole camera model has four parameters, consisting of the focal lengths and principle point. Calculating these requires 3D depth information which is not available in a single image. Classical techniques use multiple images of the same checkerboard to overcome this limitation.
This work compares various deep learning approaches to estimate the focal length of a single image. Four different methods were tested, first two methods follow an direct approach and estimate the normalized focal length or field of view. The third method lets the network estimate a 2D vector, when the last method focuses on estimating per pixel viewing rays.

Dataset

The Focalens [2] dataset was used to train the networks. Focalens [2] provides a wide variety of images with different field of view. The approximately 234,000 photos in the dataset are divided into four subsets: indoor, city, landscape, and portrait. With a roughly 2:1 ratio to portrait and indoor photos, it prioritizes city and landscape photos. Focalens [2] was specifically designed to have an equal distribution (Figure- 1) of the field of view. When plotting the distribution it showcases that the field of view covers a range from 6.88 to 121.88 with the majority of the data falling below the 60 degree mark with the mean at 48.62.

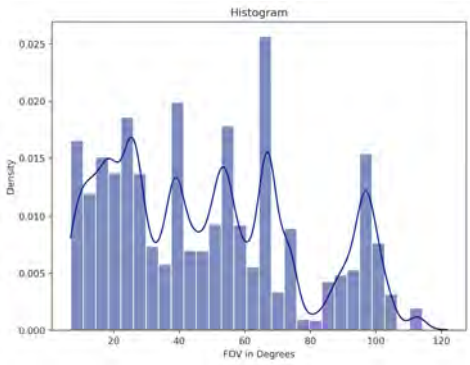


Fig. 1: Distribution of the field of view in the Focalens dataset, showing the majority below the 60 degree mark [4]

Methodology

This work wants to compare and evaluate four different approaches to estimate the focal length. Each Method is gonna be tested with two different backbones: CNN [3] and ViT [1].
Direct focal length estimation (1): A fully connected layer with a single value output is the network's head. The picture labels were computed from field of view to focal length and then normalized to the range from 0 to 1 in order to train the network. For the focal length estimation its crucial to consider the image dimension reduction caused by the network and therefore up sample the prediction of the network when comparing to the ground truth (Figure- 2).

Input size	Output size	Resulting fx
224x224	7x7	1/32

Fig. 2: Image dimension reduction of ResNet and therefore resulitng scaling factor [4]

Field of view estimation (2): A fully connected layer with a single value output is the network's head. This is used to let the network estimate the normalized (0 to 1) field of view. The field of view describes the angular extent of the visible scene and is therefore not affected by image reduction or scaling operation. As a result there is no need for further adjustments of the networks predictions (Figure- 3).

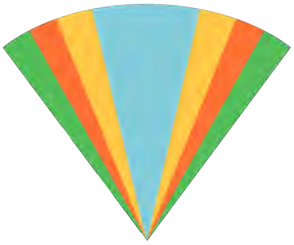


Fig. 3: Example visualization of the field of view. After normalization blue refers to 0 and green would refer to 1. The field of view describes the angular extent of the visible scene and is not affected by scaling. [4]

Indirect field of view estimation (3): The Head here is also a fully connected layer but with two instead of one output. The Network learns to estimate an 2D vector whose angle matches the field of view (Figure- 4). Since we only are interested in the direction of the vector, we use cosine similarity loss for this approach.

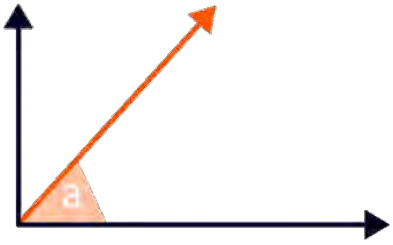


Fig. 4: Example visualization of the estimated vector and the field of view (a). The length of the vector doesn't affect the resulting angle. [4]

Viewing rays estimation (4): We let the network learn how to predict per pixel viewing rays. As a result, the Head is a 3D structure with a depth of three and the dimensions of the input image. The rays can be calculated out of the intrinsic parameters and therefore contain the information about the field of view. By using the left and right most predicted ray of a image we can calculate the field of view, which corresponds to the angle between the two rays (Figure- 5).

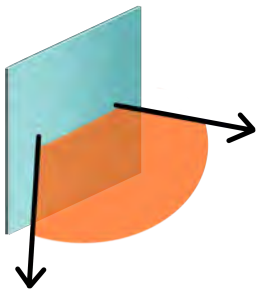


Fig. 5: Visualization of how viewing rays correspond to the field of view, where its represented by the orange area [4]

Re-projection error

The re-projection error is widely used for various 3D application as 3D Reconstruction or Pose estimation. We want to introduce the usage of the re projection error for measuring the accuracy of the estimated focal length. By using different data with known intrinsic and extrinsic parameters. We calculate the re-projection error by taking the estimated focal length and compare it to the original focal length. The error is calculated by the distance of the two pixel, displayed as the yellow line (Figure- 6). It is important to consider scaling the metric across the different images size to make it comparable.

Fig. 6: Re-projected pixels after multiplying the focal length by 0.96 at a focal length of 959px. Small changes in the focal length show an noticeable projection error, the error increases corresponding to the distance of the image center. [4]

Evaluation and Results

First test runs of the methods number (1) and (2) show that the network adapts better in estimating the field of view rather then directly estimating the focal length. Field of view is a more image-centric, intuitive, and robust parameter for neural networks to estimate, as it closely aligns with visible geometric properties of the scene and is less dependent on external camera specific metadata. Also first tests show more stable results when using an Vision Transformer for the backbone rather then a Convolutional Neural Networks. Unimportant image features as the sky or the water can be ignored, the network can focus on the feature rich parts of the scene and weigh them with the self-attention mechanism (Figure- 7). Although the ViT backbone has more stable results, it still has heavy outliers in the estimation (Figure- 8). The results can still change as the work is currently not finished.

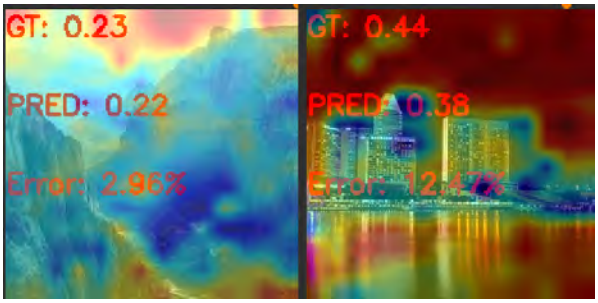


Fig. 7: Extracted attention maps (red= low, blue= high) show us where the network is focusing. The attention maps highlight that foreground structure is primarily used during inference [4]

Fig. 8: Outliers in the field of view (degrees) estimation, even though the scene doesn't change heavily. The outliers occur in both direction and showcase the stability problems in the estimations. [4]

Outlook

The current approaches don't come near the precision of classic methods as the Checkerboard-Calibration. However, they can prove valuable depending on the use case. In scenarios where a trade-off in precision is acceptable in exchange for benefits such as faster processing or the ability to work with a single image without additional metadata, deep learning becomes a viable option. Using a deep learning approach for calibration will allow us to collect and calibrate a large number of data from different sources, it will grant us the ability to work with more diverse and realistic data.

References and figures

[1] Alexey Dosovitskiy, Lucas Beyer, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv: Computer Vision and Pattern Recognition*, 2021.

[2] Yan Han, Yu Zhang, et al. Focal length estimation guided with object distribution on FocaLens dataset. *Journal of Electronic Imaging*, 2017.

[3] Keiron O'Shea, Ryan Nash, et al. An Introduction to Convolutional Neural Networks. *arXiv: Neural and Evolutionary Computing*, 2015.

[4] Own representation.

Codequalität in JavaScript-Projekten: Untersuchung und Evaluation von Tools zur Code Analyse und Optimierung

Julian Morys **Dennis Grewe**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma adesso SE, Stuttgart

Einleitung

JavaScript ist eine der weltweit am weitesten verbreitete Programmiersprachen. Hauptsächlich wird es für eine dynamische Webentwicklung verwendet. Die Entwicklung scheint allerdings dahin zu gehen, dass auch andere Aufgaben von JavaScript Skripten übernommen werden. Unter anderem findet JavaScript in Backend-Anwendungen (Node.js), mobilen Apps (React Native) und sogar Desktop-Anwendungen (Electron) Verwendung [2]. Für den steigenden Einsatz wird es nun immer wichtiger die richtigen Tools zur Qualitätssicherung zu haben.
JavaScript ist, durch weniger strikte Syntax und Logik Regeln, vergleichsweise einfach zu lernen und wird dadurch oft als Einstiegsprogrammiersprache gewählt. In professionellen Projekten ist die Qualität des Codes mitentscheidend über den Erfolg. Fehleranfälliger, beziehungsweise schwierig wartbarer Code, wird zu zusätzlichem Aufwand führen, der nicht durch Mehreinnahmen kompensiert werden kann.
JavaScript hat im Vergleich zu anderen populären Programmiersprachen, wie Java, einige Unterschiede, auf die in der Qualitätssicherung vermehrt zu achten ist, da sie zwar einen leichteren Einstieg in die Sprache ermöglichen, allerdings gerade durch ungenügende Dokumentation im Laufe des Projekts zu Problemen führen können. Eine Besonderheit JavaScripts, die in Anwendungen ohne Toolunterstützung zu schwer auffindbaren Fehlern führen kann, ist beispielsweise die *dynamische Typisierung*. Dynamische Typisierung bedeutet, dass Variablen erst während der Ausführung einen Typ (String, Integer, etc.) bekommen. JavaScripts ist dazu auch noch *schwach typisiert*. Das bedeutet, dass bei einer Operation mit mehreren Variablen, unterschiedlichem Typs, die Typen konvertiert werden, sodass eine logische Operation entsteht [3].

Zielsetzung

Es müssen die richtigen Tools gewählt werden, um die Fehler richtig einschätzen und wenn nötig angehen zu können. Hierfür sollen in dieser Arbeit JavaScript Anwendungen mit verschiedenen statischen Codeanalyse Tools getestet werden. Das sind Tools die Quellcode ohne dessen Ausführung analysieren, um Fehler, Schwächen oder Verbesserungsmöglichkeiten zu identifizieren. Es wird spezifisch auf Sicherheitslücken, Wartbarkeit und Einhaltung von Standards überprüft. Die daraus entstehenden Ergebnisse werden dann verglichen, um festzustellen, worin die Stärken und Schwächen der Tools liegen und ob es möglicherweise ein eindeutig bestes Tools gibt.
Daraus ergibt sich, dass sich diese Arbeit auf Aspekte wie die Implementierung von Best Practices und die Wartbarkeit des Codes konzentrieren wird. Das Ziel ist es, zu untersuchen, wie Tools zur Code-Analyse und Optimierung beitragen können. Die fachliche Richtigkeit oder Funktionalität der Anwendung spielt in diesem Rahmen keine zentrale Rolle.
Ebenfalls wichtig für die Evaluation zwischen den einzelnen Tools sind Usability Aspekte der Tools selber. Dazu gehören unter anderem Durchlaufzeit, Implementierungszeitpunkt oder Verständlichkeit der Berichte.

Wissenschaftliche Grundlage

Die Qualität des Quellcodes ist ein zentraler Aspekt der Softwareentwicklung und ein Indikator dafür, wie gut ein Softwareprojekt, in technischer Hinsicht entwickelt, wurde. Softwarequalität bezeichnet Eigenschaften eines Quellcodes, die sicherstellen, dass dieser sowohl funktionale, als auch nicht-funktionale Anforderungen an das Projekt erfüllt. Die ISO hat hierfür mehrere Normen für die Softwareentwicklung definiert. Die ISO Norm 9126, die schließlich von den ISO Normen 25000, speziell von der ISO 25010, abgelöst wurde, definiert einige Qualitätsmerkmale die für die Softwarequalität essenziell sind [4].

Abb. 1: Qualitätskriterien gemäß ISO 9126 [1]

Da der Schwerpunkt dieser Arbeit auf der Code-Qualität beruht, werden nicht alle Eigenschaften benötigt. Funktionalität, Benutzbarkeit und Zuverlässigkeit sind bei der Bewertung von Code-Qualität zu vernachlässigen, da sie vornehmlich funktionale Aspekte adressieren. Dennoch können einige Merkmale aus der ISO-Norm für die Bewertung der Code-Qualität herangezogen werden. Hauptsächlich für die Code-Qualität relevant sind die Änderbarkeit (engl. Changeability), die Effizienz (engl. Efficiency) und die Übertragbarkeit (engl. Portability).

Vorgehensweise

Bevor Werkzeuge zur Qualitätskontrolle in der Praxis untersucht werden können, ist es notwendig, klare Bewertungskriterien zu definieren. [5].
Hierfür soll zuerst eine umfassende Literaturrecherche die wichtigsten Aspekte der Codequalität und die spezifischen Herausforderungen im Kontext von JavaScript identifizieren. Wie in der wissenschaftlichen Grundlage beschrieben, wird hierfür auf bestehende Qualitätsstandards wie ISO 25010 Bezug genommen.
Darauf folgend wird ein Kriterienkatalog entwickelt. Dazu gehören unter anderem die Fehlererkennungsrate, die Integration in den Entwicklungsprozess und die Verständlichkeit der Ergebnisse.
Basierend auf den Kriterien werden mehrere Tools wie *ESLint*, *SonarQube*, und *Prettier* ausgewählt und in einer kontrollierten Umgebung getestet.
Die Ergebnisse der Tests werden systematisch verglichen. Dabei wird bewertet, welche Tools besonders geeignet sind, um spezifische Qualitätsaspekte wie Wartbarkeit, Sicherheit oder Effizienz zu verbessern. Es wird dazu noch untersucht, zu welchem Zeitpunkt die Tools am effektivsten eingesetzt werden können – während der Entwicklung oder später im Rahmen einer CI/CD-Pipeline.

Geplante Ergebnisse

Am Ende der Arbeit soll ein umfassender Vergleich der untersuchten Tools vorliegen, der folgende Punkte adressiert:

- Bewertung der technischen Leistungsfähigkeit: Wie gut identifizieren die Tools Fehler und Schwächen im Code?
- Praktische Einsatzmöglichkeiten: Wann und wo lassen sich die Tools am effektivsten in den Entwicklungsprozess integrieren?
- Empfehlung der besten Tools: Auf Basis der Tests wird eine Empfehlung ausgesprochen, welche Tools für verschiedene Kontexte besonders geeignet sind.
- Lücken und Verbesserungspotenzial: Identifikation von Bereichen, in denen die Tools noch unzureichend sind, sowie Vorschläge für mögliche Verbesserungen.
- Kosten-Nutzen-Verhältnis: Analyse, was der finanzielle Aufwand für die Benutzung des Tools innerhalb des Unternehmens ist

Literatur und Abbildungen

[1] Wikimedia Commons. Qualitätskriterien gemäß ISO 9126. https://commons.wikimedia.org/wiki/File:ISO_9126_quality_(de).svg, 10 2016.

[2] Parshina Maria. JAVASCRIPT BEYOND THE BROWSER, 2018.

[3] Mozilla Developer Network. JavaScript data types and data structures. https://developer.mozilla.org/en-US/docs/Web/JavaScript/Data_structures, 07 2024.

[4] Shravan Pargaonkar. Quality and Metrics in Software Quality Engineering. *Journal of Science and Technology*, 2:64–65, 2021.

[5] Hironori Washizaki, Rieko Namiki, Tomoyuki Fukuoka, Yoko Harada, and Hiroyuki Watanabe. A Framework for Measuring and Evaluating Program Source Code Quality. In *Product-Focused Software Process Improvement. Lecture Notes in Computer Science*, volume 4589, pages 284–299. PROFES 2007, 2007.

Simulation des menschlichen Fahrverhaltens zur virtuellen Absicherung automatisierter Fahrfunktionen

Aaron Mueller **Thao Dang**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

Eine wesentliche Hürde für die Einführung von autonomen Fahrfunktionen ist die Validierung dieser Systeme, welche neue Herangehensweisen erfordert. Denn die bisher eingesetzten statistischen Methoden erfordern laut Studien mehrere Milliarden Testkilometer [8]. Stattdessen ist es sinnvoll, die Funktionen mit Simulationen zu testen, und sich gezielt auf kritische Situationen zu fokussieren. Man spricht dabei von *Szenario-basiertem Testen*, wenn die Szenarien auf realen Aufzeichnungen basieren auch von *Resimulation*.
Im wesentlichen gibt es zwei Arten, um Bewegungen von Verkehrsteilnehmern in Szenarien zu beschreiben (siehe Abbildung 1). Trajektorien beschreiben den Zustand (Position, Ausrichtung, Geschwindigkeit, ...) in Abhängigkeit der Zeitpunkte, und haben den Vorteil, dass diese leicht aus realen Messungen zu extrahieren sind. Im Gegensatz dazu beschreiben Manöver das sequentielle Verhalten, typischerweise in mehreren Domänen. Ein Manöver in der Geschwindigkeits-Domäne könnte z.B. lauten: "Das Fahrzeug beschleunigt auf 60 km/h über einen Zeitraum von 10 s". Der große Vorteil von Manövern ist, dass diese parametrisiert sind und dadurch sehr einfach viele Varianten eines Szenarios generiert werden können. Durch diese Abstrahierung geht jedoch auch ein gewisser Grad an Realismus verloren, u.A. die genaue Lateralposition eines Fahrzeugs in dessen Spur.
Die Relevanz dieser Vereinfachung zeigt sich am besten anhand eines Beispiels [5]: Angenommen es soll ein Algorithmus validiert werden, welcher Spurwechsel anderer Fahrzeuge erkennen soll. Ohne die laterale Varianz würde bereits ein Algorithmus, der jede leichte Abweichung von der Spurmitte als Spurwechsel erkennt, den Test bestehen. Die Simulation ist also für die Validierung solcher Funktionen nicht realistisch genug, da Fahrzeuge in der Realität häufig etwas versetzt fahren. Ein weiteres Beispiel ist das gegenüber der Realität veränderte Sichtfeld von Sensoren, wenn Fahrzeuge in der Simulation immer nur genau hintereinander fahren.

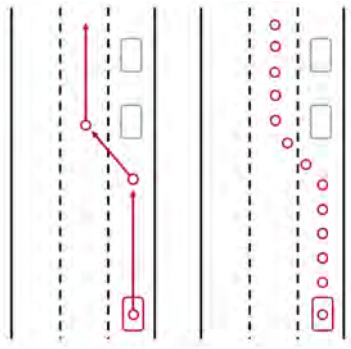


Abb. 1: Manöver-basiertes Szenario (links) verglichen mit Trajektorien-basierem Szenario (rechts) [4]

Es ist also sinnvoll, die Simulation mit dem feingranularen Verhalten der einzelnen Verkehrsteilnehmer anzureichern, um den nötigen Grad an Realismus wiederherzustellen und die Variierbarkeit der Manöverbasieren Beschreibung nutzen zu können. Das Ziel dieser Arbeit ist es, ein solches Modell zu entwickeln. Es gibt bereits mehrere Modelle, die das laterale Fahrverhalten von Menschen versuchen nachzubilden. Jedoch erfüllen diese nicht alle gewünschten Eigenschaften. Insbesondere wäre es wünschenswert, ein Modell mit erweiterten generativen Fähigkeiten zu verwenden, anstatt der bisherigen prädiktiven und autoregressiven Modelle. Deshalb soll ein alternativer Ansatz, basierend auf Deep Learning, verfolgt werden.

Anforderungen an das Verhaltensmodell

Im Vorfeld wurden einige Eigenschaften identifiziert, die ein ideales Modell erfüllen sollte.
Das Modell soll variantenreiche und realistische Sequenzen generieren, und dabei möglichst effizient sein. Realistische Sequenzen zeichnen sich durch einen plausiblen zeitlichen Verlauf sowie statistische Eigenschaften aus, die konsistent mit den realen Daten sind. Darüber hinaus ist eine flexible Länge der generierten Sequenzen wünschenswert, um diese

gezielt für verschiedene Szenarios generieren zu können. Eine einfache und effiziente Kalibrierung ist essentiell, da die Erhebung von realen Trainingsdaten aufwändig und teuer ist, besonders wenn mehrere Instanzen trainiert werden sollen um fahrerspezifisches Verhalten abzubilden. An realen Daten wurde beobachtet, dass das Lateralverhalten von externen Faktoren (z.B. Fahrzeuge auf angrenzenden Spuren) abhängt [7], sodass diese berücksichtigt werden sollen. Schließlich soll das Modell transparent sein und nachvollziehbare Sequenzen erzeugen.
Der letzte Punkt ist eine generelle Schwäche des Deep Learnings, dennoch gibt es Unterschiede zwischen verschiedenen Modellen.

Vorhandene Verhaltensmodelle

Das *Doblog* Modell von Delpiano [2] ist ein autoregressives Modell der zweiten Ordnung. Es berechnet die Lateralbeschleunigung als Funktion von Lateralposition und -geschwindigkeit und einem Zufallsterm. Der Zufallsterm setzt sich aus zwei logistischen Funktionen zusammen, welche ruhige Phasen sowie plötzliche Korrekturen nachbilden. Bei Betrachtung der einzelnen Sequenzen zeigt sich jedoch, dass das Modell nicht in der Lage ist die Form der realen Verläufe einzufangen.
Qi et al. [6] modellieren das Lateralverhalten mit einem Differentialgleichungssystem, welches Lateralposition und Rauschen definiert. Letzteres basiert auf einer Transformation von Brownschem Rauschen. Bei längeren Sequenzen tendieren durch dieses Modell generierten Offset-Profile jedoch dazu, gegen Extremwerte zu streben.
Mit dem Ziel, den Einfluss von autonomen Fahrzeugen auf die Abnutzung von Straßenbelag zu schätzen entwickelten Qi und Hu ein weiteres Modell [7]. Die Lateralbewegung wird in diesem als Funktion von anziehenden, abstoßenden und störenden Kräften beschrieben, welche individuell modelliert werden. Dadurch ist das Modell einerseits sehr flexibel, andererseits sind Modifikationen an dem Modell (z.B. das Einfügen weiterer Faktoren) mühsam und komplex.
Die realistischsten Sequenzen liefert das Modell von Neis und Beyerer [5] (siehe Abbildung 2), welches die Lateralbewegung in einen groben und einen feinen Teil unterteilt. Die grobe Bewegung wird durch eine Markov-Kette erzeugt und dann mit einem Gauß-Filter geglättet. Die feine Bewegung wird mit einem davon unabhängigen Modell aus weißem Rauschen generiert. Um externe Faktoren in das Modell zu integrieren, schlagen die Autoren vor, die Markov-Kette durch ein Hidden-Markov-Modell zu ersetzen. Die beiden Modelle wurden auf einem nicht-öffentlichem Datensatz trainiert, welcher aus je mehrere Stunden Fahrdaten von wenigen Fahrern besteht, und auch für diese Arbeit verwendet werden soll. Erwähnenswerterweise schafft es das Modell, das individuelle Verhalten der Fahrer nachbilden.

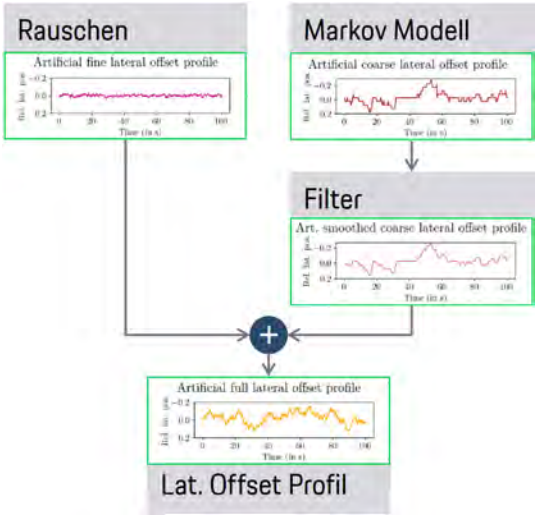


Abb. 2: Generierung von künstlichen Offset-Profilen mit dem Modell von Neis und Beyerer [5]

Überblick über generative Deep Learning-Modelle

In diesem Abschnitt werden verschiedene State of the Art Klassen von generativen Deep Learning Modellen kurz vorgestellt. Eine vollumfassende Analyse ist aufgrund der Menge an Modellen schwierig, jedoch helfen Benchmarks, wie z.B. TSGBench [1], einen Überblick über die Stärken und Schwächen einzelner Modelle zu bekommen (siehe Abbildung 3).
Autoregressive Modelle, wie **rekurrente neuronale Netze (RNNs)** oder **temporale Faltungsnetze (TCNs)**, generieren Sequenzen, indem sie schrittweise Werte aus vergangenen Informationen berechnen. Diese Modelle sind für Zeitserien besonders effektiv, da sie die sequentielle Natur der Daten direkt modellieren. Jedoch haben sie aufgrund ihrer Struktur oft Probleme, langfristige Abhängigkeiten einzufangen.
Variational Autoencoders (VAEs) lernen eine latente Repräsentation der Daten und können daraus neue Sequenzen generieren. Die probabilistische Struktur der latenten Variablen ermöglicht es, Unsicherheiten und Variationen abzubilden. VAEs haben starke generative Fähigkeiten, aber schaffen es oft nicht, die Schärfe der Trainingsdaten zu erreichen.
Generative Adversarial Networks (GANs) beschreiben ein Framework für das unüberwachte Lernen von generativen Netzwerken. Hierbei konkurrieren ein Generator und ein Diskriminator. Der Generator versucht Sequenzen zu generieren, die nicht von echten Sequenzen unterscheidbar sind, während der Diskriminator darauf traininert wird diese zu unterscheiden.

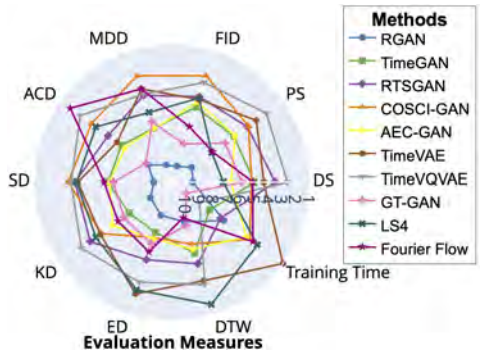


Abb. 3: Ranking der von TSGBench evaluierten Modelle in verschiedenen Tests [1]

GANs erfreuen sich aufgrund der hohen Qualität der Ergebnisse großer Beliebtheit, und sind flexibel mit anderen Architekturen, wie z.B. RNNs, kombinierbar. Eine Schwierigkeit bei GANs ist es, das Gleichgewicht zwischen Generator und Diskriminator zu finden, was in der Praxis häufig zu Instabilität führt.

Transformer-Modelle haben die Sequenzgenerierung regelrecht revolutioniert und sind die dominante Architektur der letzten Jahre in diesem Feld. Sie nutzen Self-Attention Mechanismen, um Abhängigkeiten zwischen beliebigen Positionen der Eingabesequenz parallel zu modellieren. Dies führt zu sehr realistischen Ergebnissen und herausragender Generalisierung, jedoch erfordert das Training viele Daten.

Eine sehr vielversprechende Klasse von Modellen sind **State-Space Modelle (SSMs)**. Sie modellieren den internen Zustand eines Systems mit Differenzialgleichungen. Klassische Zustandsbeschreibungen, wie sie z.B. in Kalman-Filtern verwendet werden, müssen sorgfältig modelliert werden, sowohl in der Wahl der Zustandsvariablen als auch durch die Definition der Differenzialgleichungen. Die Idee von modernen SSMs wie S4 [3], ist es, diesen Prozess zu umgehen und die resultierenden Matrizen stattdessen lernbar zu machen. In der Praxis weisen durch SSMs generierte Zeitserien hohe Qualität auf, besonders hinsichtlich Langszeitabhängigkeiten, die teilweise sogar die der Transformer übertrifft. Das ist beachtlich, da sie verglichen mit diesen auch deutlich effizienter sind.

Ausblick

Unter diesen Modellen soll nun ein geeignetes ausgewählt werden, was am besten den Anforderungen entspricht. Eine grundlegende Frage ist auch, wie das Problem formuliert werden soll: Es wäre sowohl eine schrittweise Vorhersage als auch eine Generierung der ganzen Sequenz auf einmal denkbar.

Nach Implementierung und Training des gewählten Modells sollen die generierten Sequenzen anhand qualitativer und quantitativer Metriken bewertet werden. Dadurch soll festgestellt werden, wie gut sich das Modell für die Generierung von Lateral-Offset-Profilen im Vergleich mit den bisherigen Modellen für Manöverbasierte Simulationen eignet. Gleichzeitig helfen die Ergebnisse die anderen Modelle besser zu verstehen, und können Verbesserungspotentiale aufzeigen.

Literatur und Abbildungen

[1] Yihao Ang et al. TSGBench: Time Series Generation Benchmark. In *Proceedings of the VLDB Endowment*, volume 17, pages 305–318. VLDB Endowment, 2023.

[2] Rafael Delpiano. Understanding the Lateral Dimension of Traffic: Measuring and Modeling Lane Discipline. *Transportation Research Record*, 2675:1030–1042, 2021.

[3] Albert Gu et al. Efficiently Modeling Long Sequences with Structured State Spaces. https://arxiv.org/abs/2111.00396, 08 2022.

[4] Francesco Montanari et al. Maneuver-based Resimulation of Driving Scenarios based on Real Driving Data. *2021 IEEE Intelligent Vehicles Symposium (IV)*, pages 1124–1131, 2021.

[5] Nicole Neis and Jürgen Beyerer. Efficiently Modeling Lateral Vehicle Movement Including its Temporal Interrelations Using a Two-Level Stochastic Model. *IEEE Open Journal of Intelligent Transportation Systems*, 5:566–580, 2024.

[6] Hongsheng Qi et al. Stochastic lateral noise and movement by Brownian differential models. *2022 IEEE Intelligent Vehicles Symposium (IV)*, pages 98–103, 2022.

[7] Hongsheng Qi and Xianbiao Hu. Behavioral investigation of stochastic lateral wandering patterns in mixed traffic flow. *Transportation Research Part C: Emerging Technologies*, 155, 2023.

[8] Walther Wachenfeld and Hermann Winner. Die Freigabe des autonomen Fahrens. In *Autonomes Fahren*. Springer Vieweg Berlin, Heidelberg, 1 edition, 2015.

Implementierung von Lean Portfolio Management in Unternehmen: Die Rolle von Apptio Target Process bei der digitalen Transformation von Mercedes-Benz

Christian Mumcuyan **Thomas Rodach**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Mercedes-Benz Group AG, Stuttgart Vaihingen

Motivation

Die Automobilindustrie steht im Spannungsfeld zwischen traditioneller Produktion und der fortschreitenden digitalen Transformation. Innovative Technologien wie E-Mobilität, automatisiertes Fahren und Konnektivitätsdienste erfordern neue Ansätze im IT-Management, um die notwendige Agilität und Innovationsfähigkeit sicherzustellen. 45 % der Käufer von Elektrofahrzeugen in Deutschland wärenbereit die Automarke zu wechseln um verbesserte Sprachassistenzsysteme, automatisches Bezahlen oder assistiertes Parken und Fahren zu nutzen. [2] Dies führt zu einer steigenden Relevanz der Informationstechnologie (IT) und der Annäherung der Automobilbranche an den Technologiesektor.

Hintergrund und Problemstellung

In Anbetracht der steigenden Anzahl an IT-Applikationen sowie der damit einhergehenden Erhöhung des IT-Budgets von Unternehmen gewinnt die Einführung innovativer Methoden zur Steuerung der IT-Ressourcen zunehmend an Bedeutung. Ein effizientes Management der verfügbaren Ressourcen ist eine unabdingbare Voraussetzung, um in einem von hohem Wettbewerbsdruck geprägten Marktumfeld die erforderliche Agilität und Innovationskraft zu bewahren. Das Konzept des Lean Portfolio Managements (LPM) bietet eine strukturierte Vorgehensweise, die es Unternehmen ermöglicht, IT-Initiativen auf Basis strategischer Überlegungen zu priorisieren und Ressourcen entsprechend effektiv zuzuordnen.

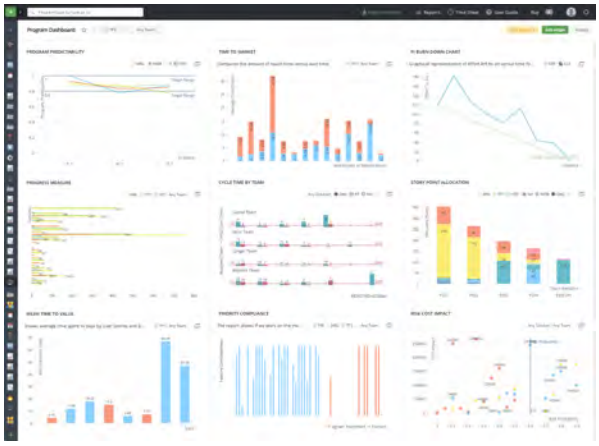


Abb. 1: Target Process [1]

Lean Portfolio Management im Überblick

LPM beruht auf drei zentralen Prinzipien: (siehe Abbildung 2)

- Strategische Themen und Portfolio Backlog: Unternehmen definieren strategische Ziele und priorisieren Projekte, die diese Ziele unterstützen.
- Lean Budgeting: Finanzmittel werden nicht auf Projekte, sondern auf Portfolios verteilt, um einen flexiblen Ressourceneinsatz zu gewährleisten.
- Portfolio-Kanban-System: Ein visuelles System, das Transparenz über den Fortschritt von Projekten schafft und Engpässe frühzeitig identifiziert.

Diese Elemente unterstützen Unternehmen dabei, Effizienz und Agilität zu steigern und gleichzeitig die Compliance in IT-Prozessen sicherzustellen.

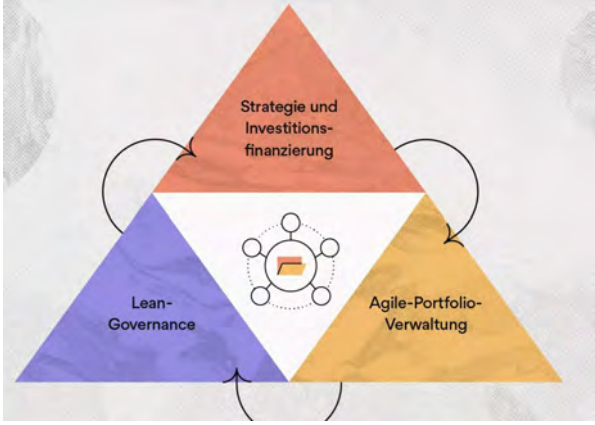


Abb. 2: LPM Dreieck [1]

Die Rolle von Apptio Targetprocess

Apptio Targetprocess ist eine flexible Enterprise-Agile-Planungslösung, die es Organisationen ermöglicht, Arbeit, Ressourcen und Portfolios dynamisch zu verwalten und dabei eine kontinuierliche Ausrichtung auf die Geschäftsstrategie sicherzustellen. Die Plattform bietet umfassende Transparenz über Projekte, Teams und Budgets, fördert die Zusammenarbeit und erleichtert die Anpassung an Veränderungen.

- Visualisierung von Portfolios: Targetprocess bietet intuitive Dashboards, die Einblicke in den Status von Projekten und deren Fortschritt ermöglichen. (siehe Abbildung 1)
- Optimierung der Ressourcenallokation: Durch automatisierte Analysefunktionen können Ressourcen effizient verteilt werden.
- KPI-Tracking: Spezifische Metriken ermöglichen eine kontinuierliche Überwachung der Portfolio-Performance und die Ausrichtung an strategischen Zielen.
- Backlog-Management: Erleichtert die Erfassung, Priorisierung und Verfeinerung von User Stories, Epics und Features innerhalb eines zentralisierten Backlogs.
- Integrationsmöglichkeiten: Nahtlose Integration mit gängigen Drittanbieteranwendungen wie Jira und GitHub, um einen reibungslosen Datenfluss zwischen verschiedenen Tools zu gewährleisten. [3]

Mercedes-Benz setzt Target Process ein, um die Effektivität von LPM zu steigern und agile Methoden auf Portfolio-Ebene zu skalieren. Die Untersuchung zeigt, dass die erfolgreiche Einführung des Tools nicht nur von technischen Faktoren abhängt, sondern auch vom Kulturwandel und der Akzeptanz durch die Mitarbeiter.

Herausforderungen und Lösungsansätze

Die Einführung von LPM in traditionellen Automobilunternehmen wie Mercedes-Benz ist mit spezifischen Herausforderungen verbunden:

- Kulturwandel: Die Akzeptanz agiler Methoden erfordert Schulungen und Change-Management-Initiativen.
- Hierarchien: Traditionelle Entscheidungsstrukturen müssen zugunsten von mehr Flexibilität aufgebrochen werden.
- Technologieintegration: Die Implementierung von Tools wie Target Process erfordert eine Anpassung der IT-Infrastruktur.

Ziele der Arbeit

- Umfassendes Verständnis der Rolle von Apptio Target Process im Unternehmen und bei der digitalen Transformation
- Prüfung, ob Apptio Targetprocess die Versprechen hinsichtlich Effizienz, Transparenz und Agilität im IT-Portfolio-Management erfüllt.
- Untersuchung der Integration und Nutzung von Apptio Targetprocess in Verbindung mit dem SAFe-Framework (Scaled Agile Framework).
- Analyse, wie Apptio Targetprocess eine konsistente und standardisierte Anwendung auf Portfolio-Ebene ermöglicht.

Ausblick

Diese Untersuchung bietet wertvolle Einblicke in die Implementierung von Lean Portfolio Management (LPM) und die Nutzung von Apptio Targetprocess als unterstützendes Tool. Die gewonnenen Ergebnisse sollen die digitale Transformation im Unternehmen gezielt vorantreiben. Darüber hinaus können die Erkenntnisse als Grundlage dienen, um Prozesse in anderen Abteilungen zu optimieren und die Skalierung agiler Methoden nachhaltig zu fördern. Mit einer umfassenden Einführung des Tools sollen sich nicht nur Transparenz und Ressourcennutzung verbessern, sondern auch die Innovationskraft und Wettbewerbsfähigkeit des Unternehmens langfristig gesteigert werden.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Silberberg Martin. Vernetzte Autos: 45% der Käufer:innen eines E-Autos würden für besseres Angebot Marke wechseln. https://www.mckinsey.de/news/presse/2024-01-05-connectivity, 2024.

[3] Rick Nucci. How to Use TargetProcess: A Comprehensive Guide. https://www.getguru.com/reference/how-to-use-targetprocess-a-comprehensive-guide#:~:text=Target-process%20is%20a%20powerful%20project,projects%20and%20other%20complex%20work., 2024.

Vision Language Models and Image Captioning on Local Machines

Tobias Naab **Markus Enzweiler**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einführung

Bei Image Captioning versucht man mithilfe eines Bildes einen präzisen Text zu diesem Bild zu generieren. Mithilfe von Vision Language Modellen (VLM) können heute automatische Beschreibungen von Bildern in natürlicher Sprache erstellt werden. Diese Arbeit untersucht hierbei die Nutzung ausgewählter Modelle auf lokalen Geräten, insbesondere auf einem Mac, um ihre Leistungsfähigkeit und Effizienz ohne Cloud Unterstützung zu evaluieren.

Motivation

VLM haben vielfältige Anwendungsmöglichkeiten. Die Nutzung auf lokalen Geräten bietet dabei Vorteile wie verbesserten Datenschutz, geringere Latenzzeiten und die Möglichkeit Offline zu agieren. Allerdings stellen die hohen Anforderungen an Rechenleistung und Speicherplatz eine große Herausforderung dar. VLMs sind sehr ressourcenintensiv, was ihren Betrieb auf Geräten mit begrenzten Kapazitäten erschwert. In diesem Kontext ist es wichtig ein Modell zu wählen, das sowohl eine gute Caption Qualität liefert, als auch Ressourcenschonend arbeitet.

Vision Language Modelle

VLMs sind fortschrittliche Large Language Modelle (LLMs), die sowohl visuelle als auch sprachliche Informationen kombinieren können, um mit Bildern zu arbeiten. Die große Präsenz von LLMs wie VLMs ist der Transformer-Architektur zu verdanken, die erstmals in dem Paper „Attention is All You Need" von Vaswani 2017 eingeführt wurde [3]. In dieser werden neue Self-Attention Mechanismen präsentiert, die komplexe Muster und Zusammenhänge in großen Datensätzen erkennen können. Dadurch können genauere Bildbeschreibungen erstellt und komplexe Szenen besser verstanden werden, was das Beantworten von Fragen zu Bildern genannt: Vision Question Answering ermöglicht.

Ein anschauliches Beispiel für die Funktionsweise von VLMs ist in Abb. 1 dargestellt, wo ein Bild beschrieben wird und eine Frage zum Bild gestellt wird.

Abb. 1: Einfaches Beispiel zur Nutzung eines Vision Language Models [1]

Methodik

Zur Untersuchung der Bildbeschreibung verschiedener Modelle auf einem Mac, wurden mehrere Schritte durchgeführt:

Modellauswahl: Es wurden mehrere VLMs [2] [4] zur Evaluation ausgewählt. Dabei wurde der aktuelle Stand der öffentlich verfügbaren Modelle untersucht, um möglichst aktuelle Modelle auszuwählen. Aufgrund diverser Kompatibilitätsprobleme mit der Mac-Architektur konnten nicht alle neuen Modelle in ihrer Grundform ausgewählt und lokal genutzt werden.

Der Grund hierfür liegt darin, dass viele VLMs für den Betrieb mit Nvidia Grafikkarten optimiert sind, hauptsächlich aufgrund der CUDA Unterstützung, die von vielen Machine Learning Bibliotheken vorausgesetzt wird. Apple Silicon verwendet eine andere GPU Architektur, was die Auswahl der Modelle einschränkte. Untersucht wurden folgende Modelle:

- LLaVA-1.5 7B/13B: Open-Source, erste Wahl für experimentelle Anwendungen.
- Florence-2 Base/Large: Hohe Detailgenauigkeit, ideal für anspruchsvolle Bildbeschreibungen.
- Moondream-2: Effizient, ressourcenschonend, schnelle Verarbeitung.

Datensätze:

Zur Evaluation der ausgewählten Modelle wurde zum einen ein standardisierter 2014 Coco Caption Datensatz ausgewählt, sowie ein eigener Datensatz erstellt, welcher Alltagssituationen darstellt. Der Aufbau der Datensätze besteht aus 5 Referenzbeschreibungen je Bild. Mit dem Coco Datensatz wurden 50 Bilder beschrieben, wohingegen der eigene Datensatz aus 20 Bildern bestand.

- COCO-Datensatz: Nutzung von Referenzbeschriftungen für metrische Analysen (50 Bilder).
- Eigener Datensatz: Erstellung eines Bilddatensatzes für die Analyse in Alltagsituationen (20 Bilder).

Evaluierung:

Alle ausgewählten Modelle werden mit Metriken wie BLEU, CIDEr und METEOR auf den Datensätzen durchlaufen. BLEU bewertet die Genauigkeit durch den Vergleich mit Referenzbeschreibungen, während METEOR auch semantische Ähnlichkeiten berücksichtigt. Außerdem wurden die jeweiligen Modelle möglichst unter denselben Bedingungen hinsichtlich ihrer Geschwindigkeit untersucht, die diese zur Erstellung einer Beschreibung benötigen. Abb. 2 zeigt hierzu die Messergebnisse der jeweiligen Modelle, wobei Florence-2 Base mit 1,4s am schnellsten performt. Eine Umfrage soll im späteren Verlauf ermitteln, welches Modell die besten Beschreibungen aus Nutzersicht erstellt.

- Evaluation: Durchführung von Automatisierter Evaluierungen mit gängigen Metriken wie BLEU, CIDEr und METEOR.
- Leistungsanalyse: Messung von Inferenzzeiten und Tokens pro Sekunde.
- Umfrage: Entwicklung einer Nutzerstudie zur zusätzlichen Bewertung der Beschreibungsqualität.

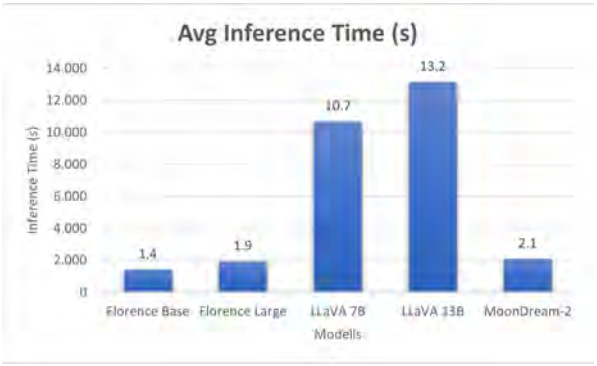


Abb. 2: Vergleich der Inferenzzeiten der getesteten VLM-Modelle auf Mac-Architektur [1]

Ausblick

Die Weiterentwicklung von VLMs eröffnet vielfältige Anwendungsmöglichkeiten in Bereichen wie Gesundheitswesen, Bildung und Robotik. Allein während dieser Arbeit, sind mehrere innovative Modelle erschienen, die durch verbesserte Weiterentwicklungen der Transformer Architektur noch präzisere Bildbeschreibungen ermöglichen. Durch weitere Forschung an der Modellarchitektur und das Training an umfangreicheren Datensätzen werden zukünftige VLMs ihre Leistungsfähigkeit und Genauigkeit weiter steigern. Gleichzeitig sind Optimierungen der Modelle und Modellgröße entscheidend, um die Effizienz und Kompatibilität auf lokalen Geräten zu ermöglichen. Trotz der Herausforderungen zeigt der Trend der letzten zwei Jahre, dass VLMs immer leistungsstärker werden und auch auf lokalen Geräten genauer sowie schneller betrieben werden können. Langfristig wäre eine breitere Nutzung von VLMs in alltäglichen Anwendungen denkbar.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] H. Liu, C. Li, Q. Wu, and Y. Lee. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems 36*. Conference on Neural Information Processing Systems, 2023.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*. Neural Information Processing Systems Foundation, 2017.

[4] B. Xiao, H. Wu, W. Xu, X. Dai, H. Hu, Y. Lu, M. Zeng, C. Liu, and L. Yuan. Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

Wie kann Künstliche Intelligenz die Prozesse in der Beschaffung unterstützen?

Kubilay Oeztopcu **Dieter Morgenroth**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Schuler Pressen GmbH, Göppingen

Die fortschreitende Anwendung von Künstlicher Intelligenz (KI) verändert das Beschaffungsmanagement grundlegend. Automatisierte Prozesse, optimierte Lieferketten und datenbasierte Entscheidungen ermöglichen Unternehmen Effizienzsteigerungen, Kostensenkungen und Risikominimierung. Angesichts komplexer globaler Lieferketten und wachsender Anforderungen wird der Einsatz moderner Technologien wie KI immer relevanter. Diese Arbeit analysiert die Potenziale von KI im Beschaffungsmanagement und zeigt praxisorientierte Ansätze für ihre Implementierung. Neben der Automatisierung von Routineaufgaben und der Optimierung von Entscheidungsprozessen werden auch Herausforderungen, wie technische und ethische Aspekte, beleuchtet. Ergänzt durch empirische Erkenntnisse und Praxisbeispiele bietet die Arbeit konkrete Handlungsempfehlungen für die Integration von KI-Technologien.

Grundlagen der Beschaffung

Die Beschaffung ist eine zentrale Funktion der Wertschöpfungskette eines Unternehmens, die sicherstellt, dass benötigte Materialien und Dienstleistungen effizient bereitgestellt werden. Traditionelle Beschaffungsprozesse umfassen mehrere Schritte, darunter Bedarfsmeldung, Genehmigung, Lieferantenauswahl, Warenprüfung und Rechnungsfreigabe. Dabei spielt der Einkauf eine entscheidende Rolle, da er Preise, Qualität und Lieferbedingungen bewertet und die Bestellungen auslöst. Diese bewährten Prozesse gelten jedoch zunehmend als zeitaufwendig und unflexibel, insbesondere in einer digitalisierten Geschäftswelt. Elektronisch gestützte Beschaffungsprozesse automatisieren viele dieser Schritte, erhöhen die Transparenz und reduzieren Fehler. Beispielsweise prüft ein digitaler Workflow automatisch das Budget und leitet nach Freigabe die Bestellung direkt an den Lieferanten weiter, was die Effizienz deutlich steigert [4].

Abb. 1: Digitaler Einkaufsprozess in Unternehmen [1]

Die Digitalisierung bildet die Basis für den Einsatz von Künstlicher Intelligenz, die durch Datenanalyse und Mustererkennung traditionelle Prozesse weiter optimieren kann. So eröffnet sich für Unternehmen die Möglichkeit, Beschaffungsvorgänge nicht nur zu modernisieren, sondern auch langfristig effizienter und agiler zu gestalten.

Künstliche Intelligenz im Überblick

Künstliche Intelligenz (KI) ist eine Technologie, die Systeme befähigt, eigenständig Probleme zu lösen und Entscheidungen zu treffen. Kern der modernen KI ist das Maschinelle Lernen, das Algorithmen ermöglicht, aus Daten zu lernen und Muster zu erkennen. Im Beschaffungsmanagement optimiert KI-Prozesse, automatisiert Routinetätigkeiten und verbessert Entscheidungen. Damit eröffnet sie Unternehmen neue Möglichkeiten, Effizienz zu steigern und Kosten zu senken [3].

Einsatz von KI in der Beschaffung

Künstliche Intelligenz (KI) automatisiert Routineaufgaben wie die Verarbeitung von Bestellungen, Datenpflege und Rechnungsprüfung. Algorithmen analysieren Daten in Echtzeit, reduzieren Fehler und

schaffen Freiräume für strategische Tätigkeiten. Ein weiteres Potenzial von KI liegt in datenbasierten Entscheidungen. KI-Systeme analysieren große Datenmengen, erkennen Muster und liefern Vorhersagen zu Marktpreisen oder Lieferengpässen. Diese Erkenntnisse ermöglichen es Unternehmen, Risiken frühzeitig zu erkennen und fundierte Entscheidungen zu treffen. In der Lieferkettenoptimierung verbessert KI die Effizienz durch Echtzeit-Analysen und Vorhersagen, etwa zu Verzögerungen oder externen Risiken. Technologien wie Predictive Analytics helfen, Engpässe zu vermeiden und Prozesse nachhaltiger zu gestalten [2]. KI ist damit nicht nur ein Werkzeug zur Effizienzsteigerung, sondern auch ein Treiber für Innovationen und die Zukunftsfähigkeit moderner Beschaffung.

Methodik der Befragung

Die Analyse der Beschaffungsprozesse wurde durch eine Kombination aus qualitativen und quantitativen Interviews durchgeführt, um sowohl detaillierte Einblicke in die operativen Abläufe als auch eine breite Meinungsbasis der beteiligten Mitarbeiter zu erhalten. Für die qualitativen Interviews wurden fünf persönliche Gespräche mit ausgewählten Mitarbeitern aus dem Beschaffungsbereich geführt. Diese Gespräche fanden im 1-zu-1-Format statt und basierten auf einem vorgefertigten Fragebogen. Ziel dieser Interviews war es, die bestehenden Prozesse im Detail zu verstehen und gezielt Problemfelder zu identifizieren. Besonderer Fokus lag dabei auf der Analyse, welche Tätigkeiten am meisten Zeit beanspruchen und welche Schritte potenziell durch KI optimiert werden könnten. Durch den strukturierten, aber offenen Ansatz konnten nicht nur bestehende Schwachstellen, sondern auch mögliche Lösungsansätze aus der Sicht der Mitarbeiter erfasst werden. Diese qualitativen Daten bieten eine wertvolle Grundlage für die Entwicklung spezifischer KI-basierter Optimierungsvorschläge, insbesondere in Prozessen, bei denen Automatisierung möglich ist. Zusätzlich wurden quantitative Interviews durchgeführt, um eine breitere Meinung der gesamten Abteilung einzuholen. Diese Methode ermöglichte es, allgemeine Trends und Meinungen zur aktuellen Prozessgestaltung und der potenziellen Einführung von KI zu erfassen. Mithilfe standardisierter Fragebögen wurden verschiedene Aspekte bewertet, darunter die wahrgenommene Effizienz der derzeitigen Prozesse, das Vertrauen in KI-Technologien und die Akzeptanz gegenüber Veränderungen im Arbeitsumfeld. Die aggregierten Ergebnisse dieser Befragungen bieten eine statistische Grundlage, um die Relevanz und das Verbesserungspotenzial einzelner Prozessschritte zu bewerten. Durch die Kombination dieser beiden Methoden konnten sowohl tiefergehende Einblicke in spezifische Herausforderungen als auch ein umfassendes Bild der allgemeinen Einstellungen und Meinungen gewonnen werden. Diese Erkenntnisse sind essenziell, um datenbasierte und praxisorientierte Empfehlungen für die Integration von KI in die Beschaffungsprozesse zu entwickeln.

Integration von KI in bestehende Systeme

Die Integration von Künstlicher Intelligenz (KI) in Systeme wie SAP erfordert sorgfältige Planung und Anpassung an bestehende Prozesse. Technologien wie Maschinelles Lernen und Natural Language Processing (NLP) ermöglichen es, große Datenmengen effizient zu analysieren und automatisierte Entscheidungen zu treffen. Beispielsweise können durch KI-Daten aufbereitet, Berichte automatisiert erstellt und Bedarfsprognosen verbessert werden. Ein zentraler Vorteil der KI-gestützten Datenanalyse ist die Fähigkeit, Trends zu erkennen und Bestellprozesse zu optimieren. Automatisierte Datenintegration verbindet Informationen aus unterschiedlichen Quellen, reduziert manuelle Tätigkeiten und erlaubt es Unternehmen, schneller auf Veränderungen zu reagieren. Mit NLP können strukturierte und unstrukturierte Daten, wie Vertragskonditionen, verarbeitet werden. Diese Daten bilden die Grundlage für Analysen, die Anomalien in Lieferketten aufdecken oder Bestellzyklen effizienter gestalten [2]. Durch die Kombination von Datenintegration und Automatisierung steigert KI die Effizienz und sichert langfristige Wettbewerbsvorteile.

Ausblick

Die Ergebnisse dieser Arbeit zeigen, dass Künstliche Intelligenz große Chancen bietet, die Beschaffungsprozesse effizienter und moderner zu gestalten. In Zukunft könnten KI-Systeme noch stärker integriert werden, um zum Beispiel Lieferengpässe frühzeitig zu erkennen oder automatisch bessere Entscheidungen bei der Lieferantenauswahl zu treffen. Dabei wird es für Unternehmen entscheidend sein, die Technologien schrittweise einzuführen und die Mitarbeitenden aktiv einzubinden. Weiterführende Studien könnten untersuchen, wie kleine und mittelständische Unternehmen KI erfolgreich nutzen können und welche speziellen Anforderungen sie dabei haben. Mit weiteren Entwicklungen wird KI ein immer wichtigerer Bestandteil der Beschaffung werden und Unternehmen helfen, in einer zunehmend komplexen Welt wettbewerbsfähig zu bleiben.

Literatur und Abbildungen

[1] d.velop AG. Einkaufsprozesse. https://www.d-velop.de/wp-content/uploads/sites/10/2023/04/uebersicht-einkaufsprozesse.png, 2023.

[2] Werner Bünnagel. *Künstliche Intelligenz und Unternehmenswissen*. Springer Gabler, 2024.

[3] Buxmann Peter and Holger Schmidt. *Künstliche Intelligenz: Mit Algorithmen zum wirtschaftlichen Erfolg*. Springer Gabler, 2 edition, 2021.

[4] Uwe Schmitz. *Grundkurs Electronic Business: Grundlagen, IT-Instrumente und Spezialgebiete*. Springer Vieweg, 1 edition, 2021.

Maßnahmen zur Systemhärtung eines industriellen Echtzeitsystems auf Basis von Embedded Linux

Ibrahim Omerhafizovic **Dominik Schoop**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Leuze electronic GmbH + Co. KG, Owen

Einleitung

Die Digitalisierung industrieller Systeme bietet Vorteile, birgt jedoch auch Sicherheitsrisiken. Die Einführung von Linux vereinfacht die Entwicklung und Wartung von Softwarelösungen für industrielle Steuerungs- und Automatisierungssysteme, erhöht jedoch die Angriffsfläche und erfordert ein solides Sicherheitskonzept. Normen wie IEC 62443 und Gesetze wie der „EU Cyber Resilience Act“ fordern eine hohe Sicherheit und Widerstandsfähigkeit.
Dies wird am Beispiel eines industriellen Sensors von Leuze deutlich, der auf einer selbst erstellten Embedded-Linux-Distribution basiert und Schwachstellen in seiner Software aufzeigt, die überwiegend in C und C++ entwickelt wurde. Diese Schwachstellen könnten es Angreifern ermöglichen, die Kontrolle über das System zu übernehmen.

Zielsetzung

Das Ziel der Arbeit ist es, die Sicherheit eines industriellen Embedded-Linux-Systems zu verbessern und seine Widerstandsfähigkeit gegenüber Cyberangriffen zu erhöhen, ohne die Leistung des Geräts zu beeinträchtigen. Durch die Implementierung eines Systemhärtungskonzepts werden nicht nur die aktuellen Sicherheitsstandards erfüllt, sondern auch zukünftige Anforderungen sowie Gesetze wie der „EU Cyber Resilience Act“ berücksichtigt. Zusätzlich zu den bereits bekannten Sicherheitslücken sollen, falls vorhanden, weitere Sicherheitslücken identifiziert und je nach ihrer Bedeutung behoben werden. Um die Sicherheitsziele zu erreichen, wird eines der beiden bekanntesten Linux-Sicherheitsframeworks verwendet: AppArmor oder SELinux.

MAC und DAC

„Discretionary Access Control”(DAC) ist ein Zugriffsmodell, das Linux- und UNIX-Systeme nutzen. DAC ist ein Modell bei dem Dateien und Prozesse einem Eigentümer zugewiesen werden, der Eigentümer kann ein Benutzer sein, eine Benutzergruppe oder auch andere Personen. Bei DAC erhalten die Nutzer die Möglichkeit, die Zugriffsrechte für eigene Dateien beliebig anzupassen. In einem DAC-Modell hat der Root-Benutzer keine Einschränkungen und kann auf alle Daten zugreifen. Ebenso ist es für den Root-Benutzer möglich, beliebige Änderungen im System vorzunehmen und auf die Dateien anderer Nutzer zuzugreifen
Bei Systemen mit „Mandatory Access Control”(MAC) hingegen, setzt die Systemadministration Zugriffsrichtlinien auf und entscheidet welche Zugriffsrechte ein Prozess erhält. Falls ein Benutzer die Berechtigungen für sein Verzeichnis verändert, wird der Zugriff durch andere Benutzer oder Prozesse trotzdem verweigert, wenn die festgelegten Zugriffsrichtlinien dies nicht erlauben. Dadurch bringen MAC-Systeme eine zusätzliche Sicherheit für das System. [1] Die beiden bekanntesten Beispiele für MAC-Systeme, die die Grundlage für diese Arbeit bilden, sind „AppArmor“ und „Security-Enhanced Linux“ (SELinux).

AppArmor

AppArmor ist ein Linux Security Module (LSM), das auf MAC basiert und einen profilbasierten Sicherheitsansatz verfolgt, um Zugriffsrechte auf Systemressourcen zu gewähren oder zu verweigern. Bei diesem Ansatz verfügt jede Anwendung über ein eigenes Sicherheitsprofil. Die Profile dienen dazu, zu bestimmen, welche Aktionen eine Anwendung ausführen darf, wie zum Beispiel, auf welche Dateien sie zugreifen kann. Die AppArmor Profile werden beim Booten in den Kernel geladen und können in zwei Modi betrieben werden: „Enforcement“ und „Complain“. Profile, die im Enforcement-Modus geladen werden, erzwingen die im Profil definierten Richtlinien. Zusätzlich werden Verstöße gegen diese Richtlinien protokolliert. Die Profile im Complain-Modus erzwingen die definierten Richtlinien nicht, sondern protokollieren

lediglich versuchte Verstöße, ohne diese zu blockieren. Zusätzlich unterscheidet AppArmor zwischen zwei Arten von Richtlinien: „Paths“ und „Capabilities“. Die „Paths“-Richtlinien legen fest, welche Dateien für eine Anwendung oder einen Prozess zugreifbar sind, während „Capabilities“ die Rechte für einen eingeschränkten Prozess definieren. [2]

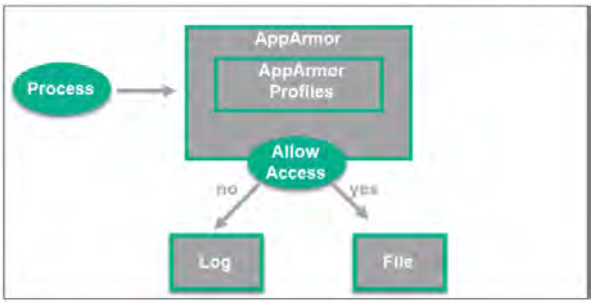


Abb. 1: AppArmor [3]

SELinux

SELinux ist eine deutlich komplexere Alternative zu AppArmor, was vor allem auf die detaillierte Definition von Zugriffskontrollen und Sicherheitsrichtlinien zurückzuführen ist. Diese Komplexität bietet jedoch auch mehr Kontrolle über das Linux-System. Wenn eine Anwendung oder ein Prozess Zugriff auf eine Datei anfragt, wird diese Anfrage zunächst von SELinux überprüft. SELinux verwendet hierfür einen Access Vector Cache, indem die Berechtigungen für die Anwendung oder den Prozess gespeichert werden. Für die Zugriffskontrolle verwendet SELinux die Konzepte „Type Enforcementünd „Labeling”. Alle Dateien, Prozesse und Ports in einem SELinux-System bekommen ein zugehöriges Kennzeichen. Die Kennzeichen dienen dazu, die Dateien, Prozesse und Ports entsprechend zu gruppieren und werden vom Kernel während des Bootvorgangs verwaltet. Durch Type Enforcement werden die festgelegten Sicherheitsrichtlinien des Systems durchgesetzt. Es legt fest, ob ein Prozess mit einem bestimmten Typ auf eine Datei zugreifen kann, die mit einem anderen Typ gekennzeichnet ist. Diese Methode sorgt dafür, dass der Zugriff auf Ressourcen strikt kontrolliert wird [1]

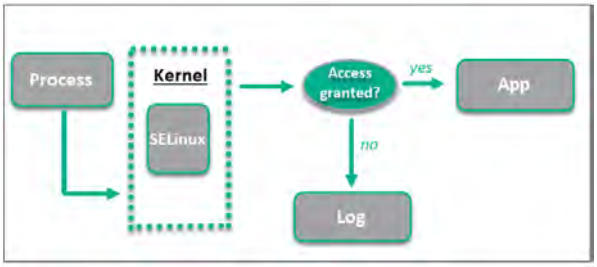


Abb. 2: SELinux [3]

Ausblick

In Zukunft wird ein Vergleich zwischen den Sicherheitsmaßnahmen AppArmor und SELinux durchgeführt, wobei besonderer Wert auf eine minimale Bootzeit und möglichst geringe Auswirkungen auf die Performance gelegt wird. Erste Informationen wurden bereits gesammelt, und es ist geplant, beide Frameworks anhand von Performanceanalysen zu testen. Ziel ist es, die geeignetste Lösung für das System zu identifizieren und dabei sowohl Sicherheitsaspekte als auch Systemeffizienz zu berücksichtigen. Eine regelmäßige Anpassung und Aktualisierung der Sicherheitsrichtlinien wird ebenfalls notwendig sein, um den steigenden Anforderungen gerecht zu werden.

Literatur und Abbildungen

[1] RedHat Incorporated. What is SELinux (Security-Enhanced Linux)? https://www.redhat.com/en/topics/linux/what-is-selinux, 08 2019.

[2] Nishit Majithia. AppArmor. https://wiki.ubuntu.com/AppArmor, 10 2024.

[3] Sara Zivanov. AppArmor vs. SELinux: Comprehensive Comparison. https://phoenixnap.com/kb/apparmor-vs-selinux, 11 2022.

Datenverwaltung für HD Map Learning

Chrysovalantis Papageorgiou **Markus Enzweiler**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Mercedes-Benz AG, Böblingen

Einleitung und Motivation

Im Bereich der Entwicklung autonomer Fahrzeuge hat sich die Nutzung von high-definition (HD) Karten als vielversprechender Ansatz ergeben. Im Gegensatz zu herkömmlichen digitalen Karten umfassen HD-Karten weitreichende Informationen der gesamten Straßeninfrastruktur auf Fahrspurebene, die zur detaillierten Wahrnehmung des Umfelds und der präzisen Lokalisierung des Fahrzeugs dienen. Sie wird durch ihre zwei Hauptkomponente beschrieben, bestehend aus einer Punktwolkenkarte und einer Vektorkarte. Da HD-Karten über das Sichtfeld der OnBoard-Sensoren hinausgehen, ermöglichen sie eine vorausschauende Bewegungsplanung und Hindernisvermeidung und unterstützen die Erstellung effizienter und sicherer Pfade [5]. Die Generierung solcher Karten hängt oft mit einer aufwendigen Datenerfassung und -aufbereitung zusammen, weshalb angenommen werden kann, dass das Einbeziehen weiterer Daten aus bestehenden Kartenquellen die Datenbasis positiv ergänzt.

OpenStreetMap

OpenStreetMap (OSM), als frei zugängliches und kollaboratives Kartenprojekt, bietet große Mengen an Geodaten die durch eine weltweite Community regelmäßig gepflegt und aktualisiert werden. Das Ziel des Projektes ist es den offenen und flexiblen Zugang der Daten bereitzustellen, welche eine vielfältige Nutzungsmöglichkeit erlauben wie z.B. für Navigation, Stadtplanung und wissenschaftlichen Forschungen. Neben den zahlreichen Vorteilen variiert die Qualität und Detailliertheit der Daten je nach Region, sodass durch die gemeinschaftsbasierte Kartenerstellung dennoch Fehler und Lücken in der Abdeckung bestimmter Gebiete vorkommen [1]. Daher ist es erforderlich, die relevanten OSM-Daten zu validieren, um deren Eignung für den spezifischen Anwendungszweck zu überprüfen und sicherzustellen, dass sie den gewünschten Nutzen bieten.

OSM-Datenmodell

Das in Abbildung 1 dargestellte OSM-Datenmodell besteht aus den drei Objekttypen Nodes, Ways und Relations, welche jeweils für die Beschreibung ihrer Eigenschaften mit Tags versehen werden können. [2]

- **Nodes** Ein geographischer Punkt, welcher üblicherweise aus einem Längen- und Breitengrad beschrieben wird. Sie definieren unter anderem Points of Interests (POIs), wie beispielsweise ein Krankenhaus, eine Ampel oder ein Straßenschild.
- **Ways** Eine Lineare Struktur oder auch Polylinie bestehend aus mindestens zwei Nodes die Straßen, Flüsse oder Radwege definieren. Ways können als geschlossene Form auftreten, bei denen der Startknoten gleich dem Endknoten ist. Dabei handelt es sich um Polygone, womit Wälder oder Seen beschrieben werden können.
- **Relations** Darstellung von logischen Zusammenhängen zwischen gleichen oder mehreren Objekttypen. Sie bestehen aus einer Liste der enthaltenen Elemente und die Beziehung, die über den Tag „type“ beschrieben wird. Beispiele für Relations können Busrouten oder administrative Grenzen sein.

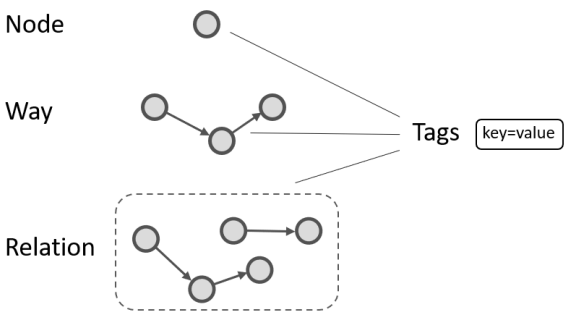


Abb. 1: OpenStreetMap Datenmodell [4]

OSM-Analyse

Für die Verwendung von OSM-Daten in Anwendungen wie der Generierung von HD-Karten, spielt die Qualität und Zuverlässigkeit der Daten eine entscheidende Rolle. Um die Genauigkeit der OSM-Daten zu evaluieren, wurden diese mithilfe einer weiteren Kartenquelle als Referenz verglichen. Die Validierung umfasst sowohl qualitative als auch quantitative Analysen, die auf bestimme Eigenschaften wie Straßensenklassen, Geometrien und der Anzahl an Spuren ausgerichtet waren. Die Ergebnisse der Analysen ermöglichten eine detaillierte Bewertung der OSM-Daten. Zudem wurden städtische als auch ländliche Regionen untersucht, um Aussagen zur Konsistenz der OSM-Daten zu treffen.

Qualitative Analyse

Um Unterschiede in der Darstellung von Straßennetzen, -klassen und geografischen Merkmalen zu identifizieren wurde zu Beginn eine visuelle Inspektion der Daten mit der Geoinfromationssoftware QGIS [3], einem Open-Source-Tool zur Analyse und Bearbeitung räumlicher Daten, durchgeführt. Die visuelle Gegenüberstellung der Datensätze ermöglichte es Diskrepanzen bei der Positionierung und Klassifizierung von Straßen oder der Vollständigkeit des Straßennetzes in spezifischen Regionen zu erkennen. Kriterien für den Vergleich:

- Präzision der Georeferenzierung: Wie exakt werden Straßen in beiden Karten dargestellt.
- Vollständigkeit: Sind bestimmte Straßensegmente enthalten oder fehlend.
- Konsistenz der Klassifikation: Stimmt die Klassenzuordnung der Straßen zwischen den Karten überein.

Die qualitative Analyse lieferte erste Einblicke in die potenziellen Stärken und Schwächen der OSM-Daten in den betrachteten Regionen.

Quantitative Analyse

Aufbauend auf den Ergebnissen des visuellen Vergleichs gegenüber einer geeigneten Referenzkarte, war Ziel dieser Analyse die Unterschiede zwischen beiden Kartenquellen hinsichtlich folgender Eigenschaften quantitativ zu bewerten, darunter:

- Übereinstimmung und Konsistenz der Straßenklassen: Welche Straßenklassen definieren eine bestimmte Straßenklasse der jeweils anderen Karte.
- Anzahl der Spuren: Analyse der Konsistenz des Attributs, welches die Anzahl der Spuren angibt.
- Geometrische Genauigkeit: Die geometrische Lage der Straßen anhand räumlicher Distanzmessungen.

Nächste Schritte

Für den weiteren Verlauf der Arbeit wird ein Encoding für die OSM-Daten entwickelt, um die Integration der Daten in eine bestehende Pipeline zur Generierung von Offline HD-Karten zu ermöglichen [6]. Das Encoding dient dazu die Merkmale der OSM-Daten in ein für die Pipeline kompatibles Format zu bringen. Anhand von Trainings sollen potenzielle Vorteile der miteinbezogenen OSM-Daten analysiert und bewertet werden.

Literatur und Abbildungen

[1] OpenStreetMap Community. About OpenStreetMap. https://wiki.openstreetmap.org/wiki/About_OpenStreetMap, 2024.

[2] OpenStreetMap Contributors. Elements. https://wiki.openstreetmap.org/wiki/Elements, 2024.

[3] QGIS Development Team. QGIS Geographic Information System. https://qgis.org, 2024.

[4] A. Jafari, A. Both, D. Singh, A. Both, and B. Giles-Corti. Building the road network for city-scale active transport simulation models. *Simulation Modelling Practice and Theory*, 114, 2021.

[5] J. Jeong, J.Y. Yoon, H. Lee, H. Darweesh, and W. Sung. Tutorial on High-Definition Map Generation for Automated Driving in Urban Environments. *Sensors*, 2022.

[6] M. Mink, T. Monninger, and S. Staab. LMT-Net: Lane Model Transformer Network for Automated HD Mapping from Sparse Vehicle Observations. *arXiv*, 2024.

Reproduktion von CAD-Modellen in skriptbasierten, parametrisierbaren Repräsentationen zum Export in unterschiedliche Zielformate

Sven Peters **Dieter Morgenroth**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma 91interactive GmbH, Filderstadt

Kurzfassung

Die 91interactive GmbH ist unter anderem im Geschäft der Systemkonfiguratoren tätig und pflegt enge Beziehungen zu verschiedenen Maschinenbau-Unternehmen. Die Besonderheit der Systemkonfiguratoren von \work-Firma ist, dass sie für ihren Fotorealismus und ihre Performanz preisprämiert sind, die Anwender ansprechen und zu dem auch noch gut dargestellt sind.
Die Verknüpfung von Computer-Aided Design (CAD) und WebGL-Technologien ist aktuell. Beispielsweise hat Bimutaike Information Tech Shanghai im Januar 2024 ein Patent auf „CAD Drawing Quality Inspection System Based on Webgl Technology“ angemeldet [3]. 91interactive GmbH bekommt häufig STEP Modelle der im Konfigurator anzuzeigenden Bauteile von ihren Auftraggebern. Allerdings können die Webapp-Systemkonfiguratoren, im Folgenden einfach Konfiguratoren genannt, aus Gründen der genannten Vision keine CAD Modelle anzeigen können. Daher ist es nötig diese Modelle in ein polygonbasiertes Dateiformat zu übertragen.
Durch Tesselation ist dies bereits möglich. Allerdings ist Tesselation mit einem gewissen Informationsverlust verbunden, da polygonbasierte Dateiformate zum Beispiel keine Rundungen abbilden können, und deshalb diese durch flache Polygone annähern.
Zum Problem wird dies dadurch, dass die Kunden der Auftraggeber von 91interactive GmbH wieder STEP Modelle der im Konfigurator geplanten Systeme benötigen. Diese von den Kunden benötigten STEP Modelle unterscheiden sich von den initialen STEP Modellen der Unternehmenskunden durch eine andere Parametrisierung [4]. Diese wird aus der Planung im Konfigurator bezogen. Zum Beispiel kann ein Bauteil eine andere Länge aufweisen.
Die Rücküberführung von polygonbasierten Dateiformaten in STEP Modelle ist ebenfalls bereits möglich, aber für den Einsatz im 3D-Druck gedacht. Wenn also beispielhaft ein kreisförmiges Rohr in der Ausgangsdatei eines Unternehmenskunden drei Meter lang ist, ein Kunde aber im Konfigurator ein zwei Meter langes Rohr plant, so erhält der Kunde durch die Tesselation und anschließende Rücküberführung eine STEP Datei von einem Rohr mit vieleckigem statt kreisförmigem Profil.
Zur Feststellung des Mehrwerts der Thesisarbeit wird die sekundäre Forschungsfrage „Kann die Performanz-getriebene Parametrisierung von 3D Modellen im Browser mit gleichzeitiger Ermöglichung des Exports in ein gleich parametrisiertes, hinreichend verlustloses CAD Format gewährleistet werden?“ gestellt.
Daher ist das Ziel der vorliegenden Arbeit, eine Repräsentation zu schaffen, welche zum einen parametrisierbar ist und zum anderen in beide gewünschten Zielformate, das STEP Modell für die Kunden und die Datei für den Konfigurator, überführt werden kann. Die Parametrisierung muss mindestens von der Datei für den Konfigurator zum STEP Modell übermittelbar sein. Die Parametrisierung kann aber auch in beide Richtungen übermittelbar sein.

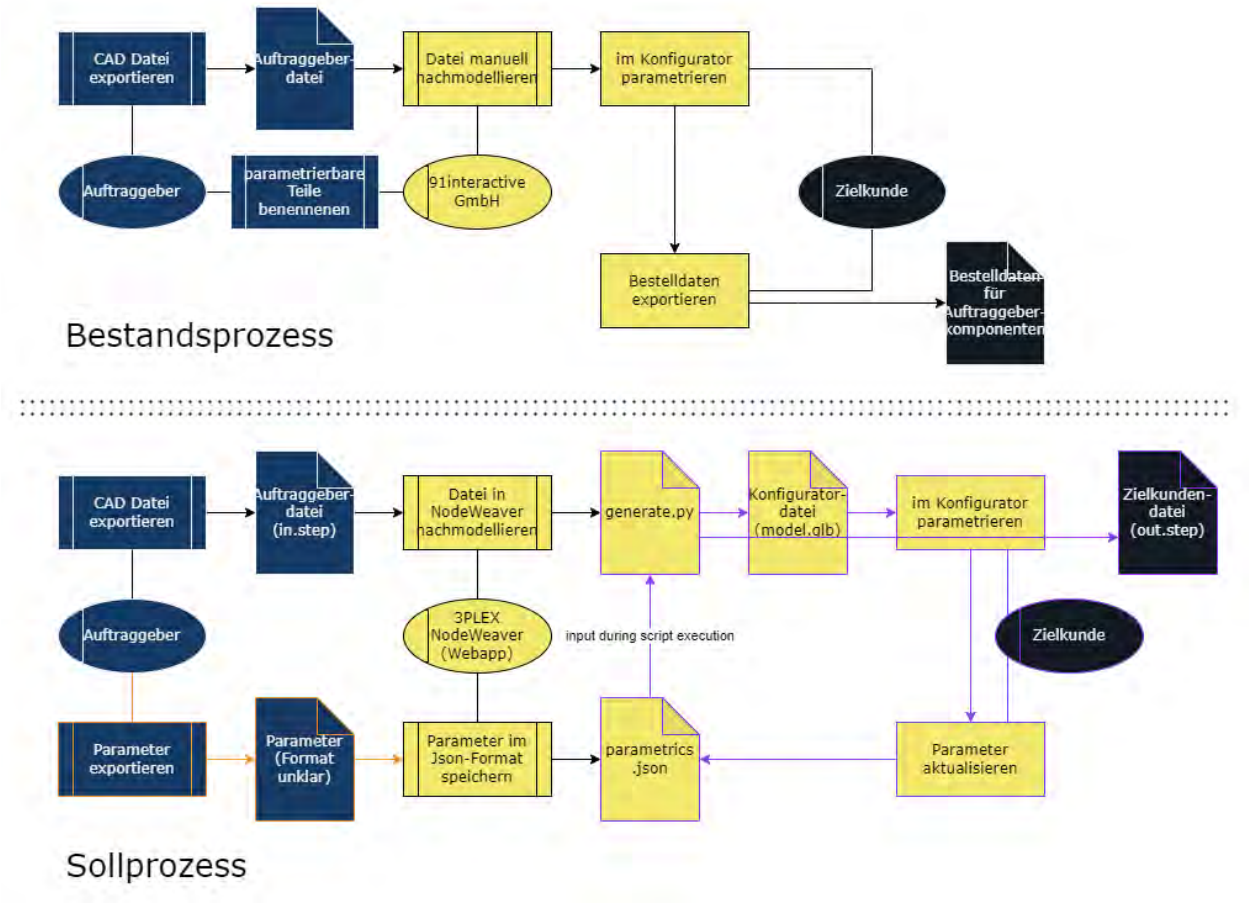


Abb. 1: Im Rahmen der Arbeit erarbeiteter Bestandsprozess und Sollprozess [1]

In Abbildung 1 sind, um dieses Ziel zu verdeutlichen, oben der vor der Arbeit implementierte Bestandsprozess und unten der nach der Bachelorarbeit zu implementierende Sollprozess mit seinen Verbesserungen gegenüber dem Bestandsprozess dargestellt. Der durch orangene Linien gekennzeichnete Unterprozess kann auch durch den „parametrierbare Teile benennen“ Unterprozess aus dem Bestandsprozess ersetzt werden, sollte dies nötig sein. Der mit lilanen Linien markierte Kreislauf ist das Ziel des Prozesses. Durch die in NodeWeaver, der nach Abschluss der Arbeit zu entwickelnden App, geschaffenen Repräsentation der Auftraggeberdatei kann ein Skript („generate.py“) erzeugt werden. Dieses Skript kann wiederum mithilfe der Parameter aus „parametrics.json“ sowohl eine glTF Datei für die Anzeige im Konfigurator, als auch eine STEP Datei für den Zielkunden generieren, letztere ist jederzeit mit der momentanen Parametrisierung aus dem Konfigurator herunterladbar. Denn werden im Konfigurator die derzeitigen Werte der Parameter aktualisiert, so werden auch die Werte in „parametrics.json“ aktualisiert und damit können sowohl die angezeigte glTF Datei als auch die Zielkundendatei jederzeit erneuert werden.

Einleitend lautet die vorrangige Forschungsfrage: „Wie kann die Parametrisierung der Längen eines Beispielmodells abgebildet werden als Repräsentation durch script-based Modelling?“. Unterstützt wird diese, neben der im obigen Paragraphen bereits genannten Forschungsfrage durch die anderen beiden Unterfragen „Wie kann diese Repräsentation in ein polygonbasiertes Format und ins STEP-Format überführt werden?“ und „Wie kann die Erlernbarkeit und Bedienung eines Webapp-Tools zur Spezifikation dieser Repräsentation für die Anwenderschaft erleichtert werden?“.

Methode

Im Rahmen der Arbeit werden, passend zu den Leitfragen, verschiedene Ansätze der Optimierung des bestehenden Prozesses zur Überführung von CAD-Modellen in unterschiedliche Zielformate, hinsichtlich ihrer Chancen und Risiken erörtert, deren Machbarkeit analysiert sowie eventuelle Schwierigkeiten und Stärken der verschiedenen Ansätze aufgezeigt.
Ziel der Arbeit ist es, die Tätigkeiten der Entwickler zu erleichtern, indem die Anforderungen, welche den bestehenden Prozessen zugrunde liegen, ver-

standen werden, der Workflow verbessert wird und eventuelle organisatorische Schwachstellen aufgezeigt werden. Außerdem wird ein Proof of Concept für die Tooling-Pipeline des Soll-Prozesses erstellt. Die Implementierung geschieht anhand einer beispielhaften Datei.
Um die Bestandsprozesse zu optimieren, müssen diese zuerst aufgenommen und analysiert werden. Nach Abschluss dieser Tätigkeiten kommt es zur Implementierung des Prototypen. Das an eine erfolgreiche operative Umsetzung anschließende Prozessreporting und Monitoring steht nicht im Umfang der Arbeit.
Für die Prozessdefinition bietet sich an, eine Umfrage mit den Anwendern des Prozesses durchzuführen. Um auch unbewusstes und nicht-bewusstes Wissen sowie unbekannte Erleichterungspotentiale eventuell aufzunehmen, wird angestrebt den bestehenden Prozess in seiner Anwendung zu beobachten.
Mit der so gewonnenen verbesserten Einsicht in die eigentliche Problemstellung können die Optionen für einen Soll-Prozess sowie dessen Toolchain erwogen werden.
Da es nach wie vor Endziel des Prozesses ist, die Parametrisierung von Modellen in webbasierten Systemkonfiguratoren zu gewährleisten, bestehen die Hauptaufgaben des Toolings voraussichtlich in Anzeige und Parametrisierung der Modelle in einer Web Anwendung.
Für die Tesselation, sollte diese notwendig werden, wird aus Gründen der regionalen Nähe zu \workFirma, insbesondere InstaLOD in Betracht gezogen.
Es wird ein Konzept für das Tooling durch agile Methoden entwickelt und in lauffähigem Proof of Concept, Code-Kommentaren und anderen geeigneten Formaten dokumentiert.

Stand der Forschung

Wie in [2] behandelt, ist es prinzipiell möglich, CAD-Dateien über die Zwischenschritte der Tesselation und verschiedener Vereinfachungen im Browser anzuzeigen. Beim Import von STL Dateien, welche derzeit für sämtliche Exports aus den Konfiguratoren von 91interactive GmbH eingesetzt werden, in CAD Programmen ist es tendenziell nicht mehr möglich diese professionell zu verändern, da sie von der jeweiligen Software in ein einzelnes rudimentäres Bauteil überführt werden. Bei STEP Dateien ist dies ähnlich, doch können diese noch in ihre Einzelteile zerlegt werden.

Zwischenstand und Ausblick

Nach initialer Evaluierung der möglicherweise einzusetzenden Frameworks, wurde ein Konzept für die Generierung des Skripts erstellt und eine Beispieldatei für dessen Implementierung ausgewählt. Die Datei musste zunächst in Cadquery nachgebildet werden um festzustellen wie die Generierung ablaufen sollte.
Es ist mittlerweile bekannt, dass die gleichzeitige Parametrisierung von STEP und glTF Modellen grundsätzlich möglich ist. Einzelheiten sind noch auszuarbeiten. Der Fokus liegt dabei insbesondere auf der Zwischenrepräsentation aus den visuellen Skript-Knoten.
Außerdem ist die effektive und effiziente Bedienung des Toolings noch zum Teil unbehandelt, womit ein überaus wichtiger Teil der Anwendererfahrung der zukünftigen Software noch nicht fest steht.
Als Zwischenfazit ist zu ziehen, dass das Ergebnis der Arbeit nur noch in dessen Details, nicht aber in seiner Machbarkeit ergründet werden wird, da letztere schon gewährleistet ist.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Helen Diez. „3D model management for e-commerce“. *Multimedia Tools & Applications 76.20*, pages 21011–21031, 2017.

[3] Global IPNews. Bimutaike Information Tech Shanghai Submits Patent Application for CAD Drawing Quality Inspection System Based on Webgl Technology. https://www.globalipnews.com/?q=Bimutaike+Information+Tech+Shanghai+, 2024.

[4] Dudenredaktion oJ. „parametrisieren“. https://www.duden.de/node/155409/revision/1321220, 2024.

Automatisierung von Arbeitsabläufen im SAP-Betrieb durch die Implementierung von Ansible

Thi Cham Pham **Gabriele Gühring**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma KfW Bankengruppe, Berlin

Einleitung

Der Begriff **Business Process Optimization** ist vielen mittlerweile vertraut. Ein Trend, der sich über verschiedene Bereiche hinweg ausbreitete und auch die IT mit einbezog, hat dazu geführt, dass im Laufe der Jahre zahlreiche Automatisierungstools entwickelt wurden. Zu den bekanntesten Tools mit einem Fokus auf Konfigurationsmanagement zählen Ansible, CFEngine, Puppet, Chef und SaltStack. Es wäre kontraproduktiv, das Hauptziel von Automatisierungstools – die Optimierung von Prozessen und die Zeitersparnis – zu verfehlen, indem man allein für das Erlernen des Tools viel Zeit aufwenden müsste, selbst wenn langfristige Effizienz einen wesentlichen Vorteil darstellt. Genau aus diesem Grund wurde Ansible entwickelt. Ein überzeugendes Argument für Ansible ist seine Benutzerfreundlichkeit: Die einfache Lesbarkeit führt zu einer geringen Lernkurve, wodurch Administratoren und Entwickler ihre Umgebung mühelos verwalten können. [6]

Konfigurationsmanagement

Konfigurationsmanagement ist eine Methode zur Verwaltung von Änderungen an Systemen, die ursprünglich in den 1950er Jahren vom US-Verteidigungsministerium eingeführt wurde. Das Ziel besteht darin, die gewünschten Zustände zu definieren, die mit Hilfe entsprechender Tools umgesetzt werden sollen, und gegebenenfalls sogar eine neue Umgebung von Grund auf zu erstellen. [4] [6] Das Konfigurationsmanagement hat nicht nur in der Serverautomatisierung an Bedeutung erlangt, sondern auch durch seinen Einsatz in Bereichen wie Virtualisierung und Containern. Wie Tim Barker bereits sagte: "*Treat your servers like cattle, not pets.*" Mit dieser Analogie soll die Transformation von der traditionellen IT, bei der Server wie unverzichtbare Einzelstücke behandelt werden, hin zur modernen Cloud-Architektur verdeutlicht werden. In dieser werden Server wie eine Herde behandelt, auf die standardisierte und automatisierte Optimierungslösungen – unter anderem durch Tools wie Ansible – angewendet werden können. [5]

Ansible

Es gibt viele weitere Gründe, sich für Ansible zu entscheiden:

- Einfachheit: Dieser Punkt wurde bereits in der Einleitung erwähnt. Zwar kann es von Vorteil sein, Kenntnisse in Python sowie in Linux- und Shell-Skripting zu haben, da diese bei der Erstellung von Modulen nützlich sein können, jedoch ist es nicht zwingend notwendig, um mit Ansible zu arbeiten, da Ansible bereits über mehr als tausend Module verfügt.
- Zustandsbasiertheit: Ansible sorgt für eine zuverlässige und wiederholbare IT-Infrastruktur, wodurch das Risiko potenzieller Fehler verringert wird.
- Sicherheit: Ansible nutzt SSH als Transportschicht zwischen Servern und Systemen, wobei die Open-Source-Software OpenSSH zum Einsatz kommt. Darüber hinaus ist Ansible agentenlos, was bedeutet, dass auf den Zielmaschinen, auf denen Befehle ausgeführt werden sollen, keine zusätzlichen Agenten installiert werden müssen, da SSH als Kommunikationsmethode verwendet wird.
- Ad-hoc-Befehle: Für einfache Aufgaben ist nicht einmal ein Playbook erforderlich. Es genügt, die Befehle zusammen mit der Zielmaschine in einer einzigen Zeile auf der Kommandozeile einzugeben und auszuführen. Auf diese Weise spart man sich sowohl das Playbook als auch ein Inventory.
- Idempotenz: Ansible führt die erforderlichen Schritte aus und erzielt immer dasselbe Ergebnis, unabhängig davon, wie oft es ausgeführt wird. [6]

Architektur von Ansible

Wie in der Abbildung 1 dargestellt, besteht die Architektur von Ansible aus sechs Komponenten:

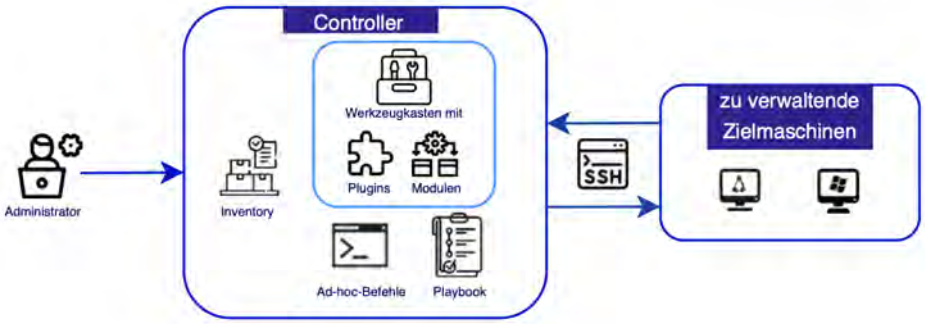


Abb. 1: Architektur von Ansible [1]

1. Module sind vorgefertigte und wiederverwendbare Automatisierungsaufgaben, die mit der Installation von Ansible bereitgestellt werden. Sie decken eine Vielzahl von Funktionen ab, wie beispielsweise das Installieren von Paketen, das Erstellen, Ändern oder Löschen von Dateien und Verzeichnissen sowie das Verwalten von Berechtigungen. Darüber hinaus ermöglichen Module das Starten, Stoppen, Aktivieren oder Deaktivieren von Diensten.

2. Ad-hoc-Befehle sind einmalige Befehle, die schnell über die Kommandozeile ausgeführt werden, wie z.B. das Abfragen des Status eines Services auf mehreren verwalteten Zielmaschinen.

3. Plugins dienen dazu, die Funktionalität von Ansible zu erweitern und komplexere Aufgaben zu ermöglichen. Dabei werden sie in die Kategorien Aktionen, Filter, Verbindungen, Callbacks und Lookups unterteilt.

4. Inventories sind Dateien, in die Benutzer Informationen über die von Ansible zu verwaltenden Zielmaschinen, wie IP-Adressen und Hostnamen, eintragen können. Dabei lassen sich die Maschinen zur übersichtlichen Verwaltung der Infrastruktur nach Funktion, Rolle, Standort oder anderen benutzerspezifischen Kriterien gruppieren.

5. Playbook ist eine Abfolge von Anweisungen, die auf den zu verwaltenden Zielmaschinen ausgeführt werden sollen. Es wird in YAML geschrieben, einer Datenserialisierungssprache, deren Name für *YAML Ain't Markup Language* steht. Dies betont die Abgrenzung zu Dokumentenauszeichnungssprachen wie XML und richtet den Fokus auf eine datenorientierte Struktur. [3]

6. Rollen ermöglichen es, den Automatisierungscode in kleinere, handhabbare Einheiten zu unterteilen, was das Testen, Debuggen und Beheben von Fehlern erleichtert. Sie können in mehreren Playbooks wiederverwendet werden. [2]

Der Administrator meldet sich am Ansible-Controller an. Anschließend generiert er einen SSH-Schlüssel und kopiert diesen auf die Zielmaschinen, um einen passwortlosen Zugriff zu ermöglichen. Danach erstellt er eine Inventardatei und wählt eine Option zur Ausführung der Aufgaben. Dabei hat er die Wahl zwischen zwei Methoden: Ad-hoc-Befehlen und Playbooks. Module und Plugins dienen als Werkzeuge für diese Methoden.

Zielsetzung

Ziel meiner Arbeit ist es, verschiedene Konfigurationsmanagement-Tools zu analysieren und miteinander zu vergleichen, dabei die Vorteile von Ansible hervorzuheben und es für die Use Cases im SAP-Betrieb innerhalb der KfW Bankengruppe zu implementieren.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Mohammed Daffalla Elradi. Ansible: A Reliable Tool for Automation. https://www.sandermann-pub.net/uploads/20230801/1bed84d124606aaf9539846a92290c1d.pdf, 08 2023.

[3] Clark Evans, Oren Ben-Kiki, and Ingy döt Net. YAML Ain't Markup Language (YAML™) Version 1.2. https://yaml.org/spec/1.1/current.pdf, 01 2005.

[4] Lorin Hochstein and Rene Moser. *Ansible: Up and Running: Automating configuration management and deployment the easy way*. O'Reilly Media, Inc., 2017.

[5] Kief Morris. *Infrastructure as Code: Managing Servers in the Cloud*. O'Reilly Media, 2015.

[6] Vincent Sesto. *Practical Ansible: Configuration Management from Start to Finish, 2*. Springer, 2 edition, 2022.

Einsatz von Machine Learning zur automatisierten Anomalieerkennung und Datenqualitätssicherung in Verkehrsdaten

Ibrahim Porsuk **Thao Dang**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

Die fortschreitende Digitalisierung und der Einsatz intelligenter Systeme im Verkehrssektor erhöhen die Anforderungen an hochwertige multimodale Daten erheblich. Sensordaten aus LiDAR, Kameras und GNSS bilden die Grundlage moderner datengetriebener Anwendungen wie präzise Objekterkennung, Umgebungsanalyse und Fahrzeugnavigation (siehe Abbildung 3). Die steigenden Datenvolumen und die Heterogenität der Quellen stellen jedoch neue Herausforderungen dar, insbesondere bei der frühzeitigen Erkennung und Behandlung von Anomalien. Ziel dieser Arbeit ist die Entwicklung eines Frameworks zur Sicherung der Datenqualität in multimodalen Verkehrsdaten.

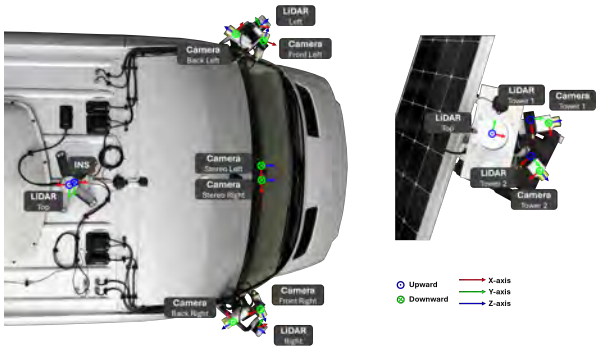


Abb. 1: LiDAR-Kamera-Projektion des Ego-Fahrzeugs [5]

Methodik

Das entwickelte Framework kombiniert regelbasierte Ansätze mit maschinellen Lernverfahren zur automatisierten Erkennung von Anomalien. Die modulare Datenpipeline wird in der Datenorchestrierungsplattform Dagster integriert [3], um datenintensive Workflows effizient und flexibel zu verarbeiten.

Maschinelle Lernverfahren

Maschinelle Lernverfahren werden speziell im Kontext der Qualitätssicherung für den Datensatz aus [5] eingesetzt, der multimodale Sensordaten umfasst, darunter Kameras, LiDAR-Daten eines Fahrzeugs sowie eines stationären Sensorturms (siehe Abbildung 1).

1. **One-Class SVM**: Diese Methode basiert auf Support Vector Machines und wird für unbalancierte Datensätze verwendet, bei denen während des Trainings nur „normale“ Daten verfügbar sind. Sie erstellt eine Entscheidungsgrenze, die die Normaldaten beschreibt, und identifiziert Datenpunkte außerhalb dieser Grenze als potenzielle Anomalien [1].

2. **Isolation Forest**: Isolation Forest ist ein Ensemble-basiertes Verfahren, das auf sukzessiven Partitionierungen des Datenraums basiert, um Anomalien effizient zu isolieren [4]. Die zugrunde liegende Idee besteht darin, dass anomale Datenpunkte weniger Partitionierungsschritte benötigen, da sie isolierter im Datenraum verteilt sind. Dieses Verfahren eignet sich besonders gut für große Datenmengen und bietet eine schnelle und skalierbare Erkennung. Abbildung 2 zeigt eine Visualisierung der mit Isolation Forest erkannten Anomalien, wobei farblich hervorgehobene Bereiche auf Datenpunkte mit hoher Wahrscheinlichkeit für Anomalien hinweisen.

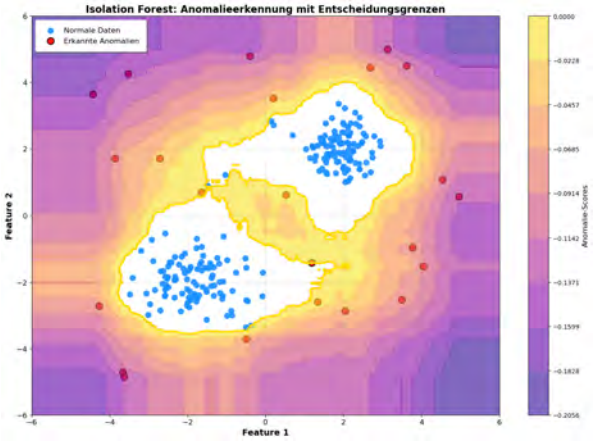


Abb. 2: Anomalieerkennung mit Isolation Forest [2]

Integration in Dagster

Um eine effiziente und reproduzierbare Verarbeitung zu gewährleisten, wurde das Framework vollständig in die Plattform Dagster integriert. Dagster ermöglicht eine flexible und skalierbare Verarbeitung multimodaler Verkehrsdaten und bietet folgende Vorteile [3]:

- Modularität: Jedes maschinelle Lernmodell und jeder regelbasierte Ansatz wird als unabhängiges Modul in Dagster implementiert. Dadurch können neue Algorithmen oder Sensordatenquellen flexibel hinzugefügt werden.
- Automatisierung: Die Verarbeitungsschritte von der Datenaufnahme über die Anomaliedetektion bis zur Ausgabe der Ergebnisse werden automatisiert und zentral orchestriert.
- Transparenz: Dagster bietet umfassende Monitoring- und Logging-Funktionen, die eine Nachverfolgbarkeit aller Arbeitsschritte gewährleisten.

Ausblick

Das entwickelte Framework hat das Potenzial, eine Lösung für die Anomalieerkennung und Datenqualitätssicherung in sicherheitskritischen Anwendungen wie autonomen Systemen zu bieten. Eine Echtzeitverarbeitung könnte dazu beitragen, Anomalien frühzeitig zu erkennen und zu behandeln, um die Effizienz und Zuverlässigkeit solcher Systeme zu verbessern.

Abb. 3: Lidar-Punktwolke mit Kameraprojektion [2]

Literatur und Abbildungen

[1] Quiroz Camilo. Erkennung von Anomalien im maschinellen Lernen: Ermitteln von Sonderfällen zur Optimierung von Geschäftsfunktionen. https://www.ibm.com/de-de/think/topics/machine-learning-for-anomaly-detection, 2024.

[2] Eigene Darstellung.

[3] Cochran Erin. Welcome to Dagster! | Dagster Docs (GitHub Repository). https://github.com/dagster-io/dagster/blob/master/docs/content/getting-started.mdx, 2024.

[4] Lorenz Maximilian and Goldstein Alexey. Anomalieerkennung mit Machine Learnings: Zwei Methoden im Fokus. https://www.convista.com/impulse/branche/versicherungswirtschaft/anomalie-erkennung-machine-learning/, 2022.

[5] Marcel Vosshans et al. LiDAR-Kamera-Projektion des Ego-Fahrzeugs. https://arxiv.org/abs/2407.08261, 2024.

Konzeptentwicklung und Implementierung einer Lösung für Cloud Diagnostic

Robert Rehberg **Rainer Keller**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Robert Bosch GmbH, Schwieberdingen

Einleitung

Moderne Fahrzeuge sind mit einer Vielzahl von Steuergeräten und Sensoren ausgestattet, die eine umfassende Analyse der Fahrzeugfunktionen ermöglichen. Die in der Steuergerätesoftware implementierten Algorithmen überwachen kontinuierlich Betriebs- und Umgebungsdaten, um Abweichungen von definierten Normalwerten zu erkennen. Diese Echtzeitüberwachung ermöglicht es, Anomalien frühzeitig zu erkennen und gezielt auf die Ursachen bestehender Probleme einzugehen, wodurch präventive und reaktive Instandhaltungsmaßnahmen optimiert werden können [4]. Der Zugriff auf Fahrzeugdaten ist jedoch nicht uneingeschränkt möglich. Es wird unterschieden zwischen gesetzlich vorgeschriebenen, standardisierten Daten, die eine Interpretation durch Drittanbieter ermöglichen, und herstellerspezifischen Daten, deren Zugriff und Nutzung häufig eingeschränkt oder komplex ist. Diese Unterscheidung beeinflusst sowohl die Diagnosetiefe als auch die Nutzbarkeit der Daten für Analysen und Anwendungen außerhalb der Herstellerumgebung.

Motivation

Die Beurteilung des Fahrzeugzustands erfolgt in der Regel durch eine Sichtprüfung. Im Vordergrund steht dabei häufig die Überprüfung der Karosserie und der Grundfunktionen wie Scheinwerfer, Klimaanlage und anderer Systeme. Zusätzlich wird bei Probefahrten auf ein normales Fahrverhalten und eine unauffällige Geräuschentwicklung geachtet. Was dabei oft vernachlässigt wird, sind die internen Fahrzeugdaten. Diese können jedoch entscheidende Informationen über den tatsächlichen Zustand des Fahrzeuges liefern, die bei herkömmlichen Prüfungen unentdeckt bleiben. Ein anschauliches Beispiel ist der Gebrauchtwagenkauf, bei dem beispielsweise der Kilometerstand eine zentrale Rolle spielt. Laut ADAC ist bei etwa jedem dritten Fahrzeug davon auszugehen, dass der Kilometerstand manipuliert wurde [1]. Solche Manipulationen können durch eine äußere Prüfung nicht zuverlässig erkannt werden. Nur die Auswertung fahrzeuginterner Daten könnte solche Auffälligkeiten aufdecken. Diese zusätzlichen Informationen verdeutlichen, wie wichtig es ist, auch das Innere eines Fahrzeugs in die Bewertung einzubeziehen. Eine fundierte Analyse der Fahrzeugdaten ermöglicht eine umfassendere Bewertung, das Erkennen von Manipulationen und fundierte Entscheidungen - sei es beim Gebrauchtwagenkauf oder bei der allgemeinen Fahrzeugdiagnose.

Zielsetzung

Das Ziel dieser Arbeit ist die Konzeption und Implementierung eines Diagnoseservices, der Anomalien und Probleme im Fahrzeuginneren auf Basis von Fahrzeugdaten erkennen und diagnostizieren kann. Dabei sollen die ermittelten Informationen so aufbereitet werden, dass sie für Nutzer ohne tiefgehende Fachkenntnisse leicht zugänglich und verständlich sind.

Perspektiven der Fahrzeugbewertung

Fahrzeuge können aus verschiedenen Perspektiven bewertet werden, je nach den Anforderungen und Zielen der beteiligten Akteure. Seien es Käufer, Verkäufer oder Organisationen wie Versicherungen. Alle diese Gruppen benötigen zuverlässige Informationen über den Zustand eines Fahrzeugs, um fundierte Entscheidungen treffen zu können. Eine zentrale Perspektive ist die des Käufers. Wie in Abbildung 1 dargestellt, gibt es beim Fahrzeugkauf klare Kriterien, die für die Entscheidungsfindung ausschlaggebend sind. Werden diese Kriterien mit auslesbaren Fahrzeugdaten kombiniert, wird deutlich, dass viele dieser Kriterien abgedeckt werden können. Beispielsweise könnte die Aufdeckung einer Manipulation des Kilometerstandes ein entscheidendes Kaufkriterium erfüllen. Eine solche Information könnte dazu führen, dass der Kauf in Frage gestellt oder der Preis erheblich nachverhandelt wird. Auch aus Sicht des Verkäufers bietet eine transparente Fahrzeugbewertung Vorteile. Ein detaillierter Bericht,

der belegt, dass keine Manipulationen oder versteckte Mängel vorliegen, stärkt das Vertrauen potenzieller Käufer. Dies kann nicht nur den Verkaufsprozess beschleunigen, sondern auch einen höheren Preis rechtfertigen. Auch Organisationen wie der TÜV Süd unterstreichen die Relevanz solcher Nachweise. Fahrzeuge mit einem Gebrauchtwagen-Zertifikat erzielen oft stabilere Verkaufspreise und werden schneller verkauft. Gleichzeitig bieten diese Zertifikate den Käufern eine wertvolle Orientierung und Entscheidungssicherheit [5]. Diese unterschiedlichen Perspektiven verdeutlichen, dass nicht nur die äußeren Merkmale eines Fahrzeugs relevant sind, sondern auch Daten, die in der Regel nicht unmittelbar ersichtlich sind. Dies unterstreicht die Notwendigkeit einer zuverlässigen internen Fahrzeugbewertung, um eine solide Grundlage für fundierte Entscheidungen zu schaffen.

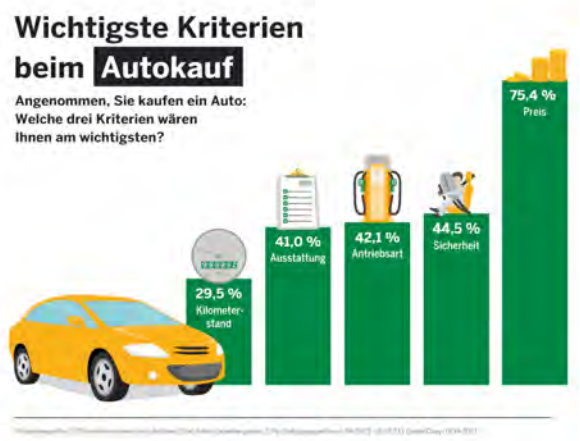


Abb. 1: Wichtige Kaufkriterien beim Autokauf [3]

Ablauf einer Fahrzeugdiagnose

Der Ablauf einer Fahrzeugdiagnose umfasst mehrere Schritte, wie in Abbildung 2 dargestellt. Im ersten Schritt wird die Auslesehardware mit der OBD2-Schnittstelle des Fahrzeugs verbunden, und die Zündung wird eingeschaltet. Anschließend aktiviert der Nutzer im zweiten Schritt die Fahrzeugdiagnose über die Benutzeroberfläche.

Im dritten Schritt empfängt die Auslesehardware ein Startsignal über das Mobilfunknetz und beginnt mit dem Auslesen aller relevanten Fahrzeugdaten. Diese Daten werden im vierten Schritt über das Mobilfunknetz an die Bosch Cloud Diagnostic gesendet. Dort werden die Fahrzeugdaten im fünften Schritt mithilfe speziell entwickelter Algorithmen von Bosch ausgewertet, um den Diagnosestatus des Fahrzeugs zu ermitteln.
Im sechsten und letzten Schritt werden die Ergebnisse der Diagnose in der Benutzeroberfläche angezeigt. Zusätzlich wird ein ausführlicher Bericht erstellt, der alle relevanten und diagnostizierten Fahrzeugdaten für weitere Nutzung bereitstellt.

Abb. 2: Allgemeiner Ablauf einer Fahrzeugdiagnose [2]

Ausblick

Je tiefer und umfassender Fahrzeugdiagnosen angeboten werden, desto größer ist der Nutzen für alle Beteiligten. Eine gründliche Diagnose ermöglicht es, nicht nur offensichtliche, sondern auch versteckte Probleme zu erkennen und transparent darzustellen. Käufer profitieren von einer besseren Entscheidungsgrundlage, während Verkäufer durch detaillierte und verlässliche Berichte das Vertrauen in ihre Angebote stärken können.
Die Integration solcher Diagnosen in cloudbasierte Plattformen bietet zudem enorme Vorteile. Daten können nahezu in Echtzeit verarbeitet, analysiert und intuitiv zur Verfügung gestellt werden. Dies erleichtert den Zugang zu komplexen Fahrzeuginformationen und macht sie auch für Laien verständlich.

Literatur und Abbildungen

[1] ADAC ADAC. Auto gebraucht kaufen: Worauf Sie achten sollten. https://www.adac.de/rund-ums-fahrzeug/auto-kaufen-verkaufen/gebrauchtwagenkauf/gebrauchtwagen-kaufen/, 2024.

[2] Eigene Darstellung.

[3] DEVK DEVK. Wichtige Kriterien beim Autokauf. https://medien.devk.de/assets/content/pressemitteilungen/pm2023/e-autos/devk-pm-2023-11-07-e-autos-umfrage-grafik-civey-und-devk-d.zip?versionId=i83HYeEt_kI101DWB4v4VaiPbfGqvgBf, 2023.

[4] Andreas Heinz. *Nutzung der Fahrzeug-Schnittstelle zur Datenerfassung im dynamischen Fahrzeug-Betrieb*. Springer Fachmedien Wiesbaden, 2024.

[5] TÜV TÜV. Gebrauchtwagen-Zertifikat. https://www.tuvsud.com/de-de/branchen/mobilitaet-und-automotive/autohaus-und-werkstatt/gutachten/gebrauchtwagen-zertifikat, 2024.

The Analysis of Homomorphically Encrypted Location Data

Denis Roth **Dominik Schoop**

Department of Computer Science and Engineering, Esslingen University

Work carried out at P3 automotive GmbH, Stuttgart

Introduction

The Global Positioning System, or GPS, forms the basis of a broad range of technological applications used in daily life. Among such applications, location-based services hold an important place, since the functionality of these applications largely depends on the user's location. These applications take the location information to enable or disable access to some service or to make recommendations based on proximity (e.g. hotel listings). For location-based services to work, the users have to share their GPS coordinates with the service provider. The provider then performs a number of tasks, such as distance calculations, based on the spatial relationship present between the shared GPS coordinates. However, sharing the precise location of the users with the server raises significant privacy concerns, particularly when the information is stored in the server's database. Privacy-related concerns are one of the major issues because people generally express apprehension about any service that may demand the disclosure of their location [1]. The location of a mobile device reveals detailed information about the movements of an individual, which can then be used for behavioral analysis, targeted advertisement, monitoring, and profiling [2]. When such sensitive information is stored in cloud infrastructures, the chances of data misuse become very high [2]. There are several ways to handle location information in a privacy-preserving manner. One method is location obfuscation. For instance, Bohli et al. [2] proposed a system that divides maps into tiles, which can be permuted and rotated by using secret keys in order to hide certain locations. Along the same lines, Kirkpatrick et al. [3] proposed an architecture that relies on the presence of a location authority in charge of issuing location verification tokens. Such tokens can be given to the service providers without having to disclose the exact location information. Another way to overcome these challenges is Homomorphic Encryption (HE). HE has proven to be an effective cryptographic method that enables computations with encrypted data. Unlike the traditional client-server model-also known as a two-party computation-HE offers a unique advantage, where the processing of encrypted data is permissible without the decryption process. This means the client provides encrypted data for secret computations and receives the results without exposing his sensitive information, since all operations are executed within the encrypted domain.

Aim of this Thesis

This thesis aims to introduce the design and implementation of a privacy-preserving location-based service utilizing homomorphic encryption. The service allows users to receive suggestions for nearby points of interest (POIs) without disclosing their actual location. It eliminates the need for trusted third parties by relying solely on a client-server architecture and on a trusted technology.

Homomorphic Encryption

In cryptography, Homomorphic Encryption (HE) is a form of encryption that enables a third-party, such as cloud providers, to directly perform operations on encrypted data, preserving at all times its structure and properties (shown in Figure 1). The concept was first introduced in 1978 by Rivest, Adleman, and Dertouzos [5], who described a theoretical framework for performing computations on data that had been encrypted. Although the concept was introduced, no practical methods were developed at that time. In Homomorphic Encryption the scheme can either be symmetric or asymmetric. What sets Homomorphic Encryption apart from the well known public-key encryption is that the algorithms for key generation, encryption and decryption rely on homomorphic mappings. That is why operations, such as addition, multiplication or subtraction are possible on ciphertext level.

Fig. 1: Processing of homomorphically encrypted data [4]

Design of a Privacy-Preserving Location-Based Service

Against the background just mentioned, the aim is to design and implement a proof-of-concept with the following requirements:

1. Only the user has access to their location data.
2. The involvement of a trusted third party is optional, so that all calculations can be outsourced to a non-trusted server.
3. Use of a generic encryption method. (There are security guidelines for encryption parameters. Such scheme also has ready-made software libraries)
4. Secure from honest but curious parties who try to learn something from the information provided but do not have the right to do so.

In this setup, the client handles location encryption and result decryption, while the server performs the required computations. The solution uses the encryption algorithms of the OpenFHE software library to minimize the potential for undetected security vulnerabilities. To address the operational constraints of homomorphic encryption, the calculation process is divided into multiple computation and communication phases. Ultimately, the client obtains personalized POI recommendations based on its encrypted request. A combination of the CKKS scheme and the FHEW scheme is used to perform the calculations. The CKKS scheme is particularly efficient compared to other schemes. It enables homomorphic operations via approximations of real numbers and thus the handling of floating point numbers. This is particularly necessary when calculating with distance metrics. In addition, the FHEW scheme, which works with logic gates and returns boolean values as results, is used to perform comparisons of homomorphically encoded floating point numbers. This schema is used to determine whether a calculated distance is within the requested radius or not.

Outlook

The results so far show that the implementation of HE in location-based services is already delivering promising results. If the scope is extended to sending homomorphically encrypted queries, the application could include not only the calculation of homomorphically encrypted locations, but also the sending of queries to an encrypted database of locations to achieve more privacy. The proof-of-concept is not yet ready for real-life application for several reasons. One reason is that a static data set of coordinates is used in the context of this work, which does not correspond to the dynamics of real-time applications. In terms of performance, the solution has a certain complexity, which is acceptable for a proof-of-concept but can still be improved. One possible way to improve this is to implement batch processing to parallelize the calculation if a certain amount of data is available. Finally, there are limitations in the cross-platform implementation, as OpenFHE only allows implementations in the programming languages Python and C++.

References and figures

[1] Louise Barkhuus and Anind K. Dey. Location-Based Services for Mobile Telephony: a Study of Users' Privacy Concerns. In *IFIP TC13 International Conference on Human-Computer Interaction*. IFIP TC13, 2003.

[2] Jens Mathias Bohli, Dan Dobre, Ghassan O. Karame, and Wenting Li. PrivLoc: Preventing Location Tracking in Geofencing Services. In *Trust and Trustworthy Computing*. Springer International Publishing, 2014.

[3] Michael S. Kirkpatrick, Gabriel Ghinita, and Elisa Bertino. Privacy-Preserving Enforcement of Spatially Aware RBAC. In *IEEE Transactions on Dependable and Secure Computing*, pages 627–640. IEEE, 2012.

[4] Own representation.

[5] Ronald L. Rivest and Michael L. Dertouzos. *On Data Banks and Privacy Homomorphisms*. No publisher, 1978.

Entwicklung eines KI-Literacy-Assistenten zur Unterstützung von Entwicklern bei der Einhaltung von CarAI Safety Anforderungen

Niko Rudolph **Thao Dang**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Mercedes-Benz AG, Böblingen

Einleitung

Künstliche Intelligenz (KI) spielt eine zunehmend bedeutende Rolle in der Gesellschaft, der Wirtschaft sowie in der Automobilindustrie. Angesichts der fortgeschrittenen KI-Entwicklung stellt sich die Frage, wie KI-Systeme sicher und gesetzeskonform entwickelt und in Fahrzeugen implementiert werden können. Durch den im Jahr 2024 erlassenen EU AI Act wird erstmals der Umgang mit KI-Systemen reguliert. Die Fülle an regulativen und internen Anforderungen macht es dem einzelnen Entwickler nahezu unmöglich, den Überblick zu bewahren. Die klassische Methode der Pflichtschulungen erweist sich aus diesem Grund als unzureichend. Einerseits ist der Zeitaufwand dieser Schulungen erheblich, andererseits finden sie selten genau zu dem Zeitpunkt statt, zu dem der Inhalt benötigt wird. Hieraus ergibt sich die Motivation, im Rahmen dieser Arbeit einen anderen Ansatz zu wählen und zu erforschen.

Motivation und Ziele

KI-Anwendungen und -Systeme müssen ebenso sicher und zuverlässig sein wie bestehende Lösungen, die auf traditionellen Algorithmen oder Hardware basieren. Insbesondere bei Hochrisiko-KI-Systemen und -Anwendungen ist diese Sicherheit von größter Bedeutung. Um dies zu gewährleisten, soll ein Ansatz verfolgt werden, der daraus besteht, wesentliche Informationen und Anforderungen zum verantwortlichen Umgang mit KI den Entwicklern bedarfsorientiert zur Verfügung zu stellen. Risikoevaluation, Maßnahmengenerierung und Compliance Checks werden in den etablierten Entwicklungsprozessen verankert. Die Entwickler sollen einen „KI-Literacy-Assistenten“ an die Seite gestellt bekommen, der sie durch diese Prozessschritte geleitet. Ziel ist, dass die Entwickler immer nur die Information erhalten, die sie situationsspezifisch benötigen.

Forschungsfragen

- Inwieweit kann ein KI-Literacy-Assistent entwickelt werden, der bedarfsorientierte und nachhaltige Unterstützung für Entwickler bietet, und welches Potenzial besitzt dieser für die zukünftige Entwicklung?
- Welche Technologien sind für die Entwicklung eines solchen KI-Literacy-Assistenten geeignet?

Theoretischer Hintergrund

Die Fortschritte der letzten Jahre im Bereich der Künstlichen Intelligenz bilden die Grundlage der Entwicklung des in dieser Arbeit angestrebten KI-Literacy-Assistenten. Um mit diesem Assistenten auf einer natürlichen Art und Weise kommunizieren zu können, wird ein Large Language Modell (LLM) benötigt. LLMs haben sich als fortschrittliche künstliche Intelligenzsysteme etabliert, die in der Lage sind, Texte zu verarbeiten und zu generieren, um eine kohärente Kommunikation zu ermöglichen und verschiedene textbasierte Aufgaben zu lösen [3]. LLMs zeigen durch ihre hunderte Milliarden vortrainierten Parameter beeindruckende Leistungen bei Aufgaben, die umfassendes Wissen über Fakten erfordern. Ihre Leistung lässt jedoch nach, wenn es um spezifischere Themen geht, bei denen keine großen Datenmengen zur Verfügung stehen [5]. Das bedeutet, dass die generierten Ausgaben Fehler oder Falschaussagen enthalten können, die den Benutzer verwirren oder seine Entscheidungen fälschlicherweise beeinflussen könne. Um diese Problematik der sogenannten "Halluzinationen"[1] gezielt zu adressieren, kann die Technik der Retrieval Augmented Generation (RAG) genutzt werden. RAG ist eine Technik zur Leistungssteigerung von LLMs durch die Integration von Informationsabrufmethoden. Halluzinationen können hierbei reduziert werden [4], indem es gezielt Informationen und Fakten

aus externen Wissensquellen abruft und als Kontext für die Generierung nutzt (siehe Abb. 1). Dadurch ist es möglich, dass RAG nicht nur faktisch korrekte, sondern auch aktuelle Antworten generiert [6]. Die enge Verzahnung von Retrieval und Generierung erlaubt es RAG, die Präzision und Aktualität der Antworten deutlich zu verbessern.

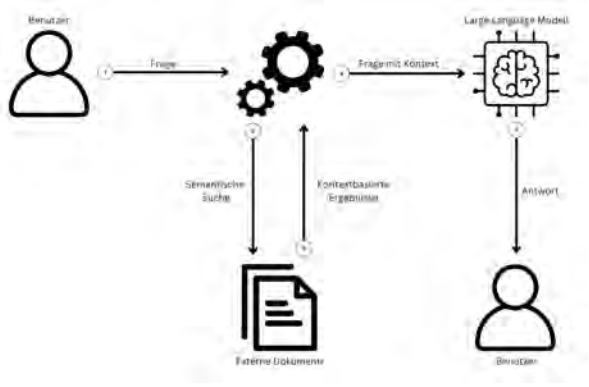


Abb. 1: RAG Architektur [2]

Ansatz und Herausforderungen

Im Rahmen der technischen Realisierung des KI-Literacy-Assistenten wurde als grundlegende Architektur das Framework Haystack gewählt, da es über die geeigneten Funktionalitäten verfügt, um eine effiziente RAG-Pipeline aufzubauen und zu erweitern. Die Pipeline-Struktur ermöglicht den unkomplizierten Einbau benötigter Komponenten sowie die mühelose Erweiterung einer bestehenden Pipeline, ohne dass eine aufwendige Konfiguration erforderlich ist. Abgesehen von der zuvor beschriebenen Pipeline basierte der bisherige Ansatz auf einer Unterteilung des Assistenten in drei Hauptaufgaben. Im ersten Schritt erfolgt die Konvertierung der zur Verfügung gestellten PDF-Dokumente in Text, gefolgt von der Reinigung, Aufteilung in Chunks, Embedding und Speicherung der Embeddings in einem Dokumentenspeicher. Im zweiten Schritt erfolgt die Retrieval-Phase, in der ein Retriever die zu einer vom Benutzer gestellten Frage relevanten Dokumente aus einem Document Store extrahiert. Die so gewonnenen Daten werden schließlich einem Generator zur Verfügung gestellt, welcher auf Basis dieser Informationen eine auf Fakten basierte Antwort generiert. Im finalen Schritt wird dem Nutzer eine Chat-Oberfläche bereitgestellt, welche eine bekannte Kommunikation mit dem KI-Literacy-Assistenten sowie eine Erinnerungsfunktion ermöglicht. Die beschriebene Architektur (siehe Abb. 2) sowie das Zusammenspiel der drei Hauptaufgaben ermöglichen die Entwicklung eines KI-Literacy-Assistenten, der über spezifisches Wissen verfügt und aktuelle Informationen faktisch richtig und mit zugehörigen Quellenangaben bereitstellen kann.

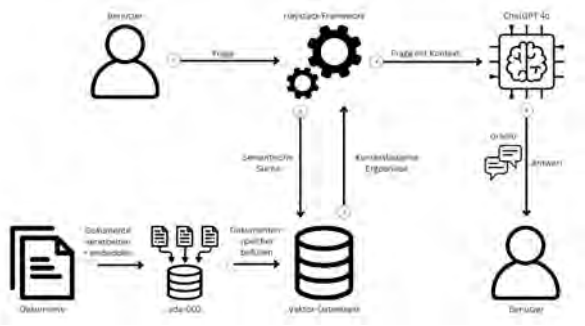


Abb. 2: AI-Literacy Assistent [2]

Die Auswahl adäquater Embedder und Generatoren stellte eine wesentliche Herausforderung dar und hatte einen entscheidenden Einfluss auf den Projekterfolg. Zu Beginn der Arbeit erfolgte die Auswahl der zu verwendenden Embedder und Generatoren auf Basis der Modellbibliothek Hugging Face. Dabei wurde eine lokale Betriebsweise angestrebt. Die Problematik bestand hierbei jedoch in der limitierten Rechenleistung sowie der Beschränkung von Modellen hinsichtlich der Verarbeitung von Tokens. Um diese Problematiken zu umgehen und das gesamte Potenzial der Pipeline auszuschöpfen, wurde die Entscheidung getroffen, als Embedder Ada-002 und als Generator ChatGPT 4o von OpenAI zu verwenden. Die Evaluierung der genannten Embedder und Generatoren hat ergeben, dass sie erstaunliche Leistungen in ihren jeweiligen Bereichen aufweisen und somit die perfekte Grundlage für den angestrebten Assistenten bieten.

Ausblick

Der aktuell erste Prototyp des AI-Literacy-Assistenten liefert bereits sehr vielversprechende Ergebnisse und lässt sich in diesem Stadium schon unternehmensintern als Prototyp vorstellen. In den nächsten Schritten liegt das Ziel darin, automatisierte Evaluierungsprozesse aufzubauen. Diese sollen die Qualität des Kontextes und die Qualität der endgültigen Antwort prüfen und bewerten. Diese Evaluierung ermöglicht es weitere Verbesserungen und Änderungen an der Architektur, den Komponenten oder den Modellen vorzunehmen und diese miteinander zu vergleichen. Damit wird eine Grundlage geschaffen, um den Prototyp iterativ zu verbessern und langfristig zu einem zuverlässigen und effektiven Werkzeug für den praktischen Einsatz weiterzuentwickeln.

Literatur und Abbildungen

[1] Y Bang, S Cahyawijaya, N Lee, W Dai, D Su, B Wilie, H Lovenia, Z Ji, T Yu, W Chung, Q V Do, Y Xu, and P Fung. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. https://arxiv.org/abs/2302.04023, 11 2023.

[2] Eigene Darstellung.

[3] H Naveed, A Khan, S Qiu, M Saqib, S Anwar, M Usman, N Akhtar, N Barnes, and A Mian. A comprehensive overview of large language models. https://arxiv.org/abs/2307.06435, 10 2024.

[4] K Shuster, S Poff, M Chen, D Kiela, and J Weston. Retrieval Augmentation Reduces Hallucination in Conversation. https://arxiv.org/abs/2104.07567, 04 2021.

[5] H Soudani, E Kanoulas, and F Hasibi. Fine tuning vs. retrieval augmented generation for less popular knowledge. https://arxiv.org/abs/2403.01432, 12 2024.

[6] H Yu, A Gan, K Zhang, S Tong, Q Liu, and Z Liu. Evaluation of Retrieval-Augmented Generation: A Survey. https://arxiv.org/abs/2405.07437, 07 2024.

Automatisiertes End-to-End-Softwaretesting anhand einer Hochschul-Website

Markus Rumpel **Andreas Rößler**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

In der heutigen Welt gibt es immer mehr und immer größere Softwareprojekte. Wo früher Programmcode im Umfang von mehreren Megabytes üblich war, ist die Menschheit mittlerweile an einem Punkt angekommen, an dem die Code - Menge bis in den Gigabytebereich reicht. Frameworks für verschiedene Zwecke werden mit der Zeit immer größer und komplexer, wodurch sich allgemein die Komplexität der virtuellen Welt erhöht. Mit zunehmender Kapazität an Speicher steigt folgerichtig auch die Anzahl der Zeilen an Quellcode. Die visuelle Überprüfung von Programmcode durch einen Menschen ist extrem zeitaufwändig und fehleranfällig, und letztlich hat jede Person nur eine begrenzte Zeit von 24 Stunden pro Tag zur Verfügung. Wenn jedoch Tests unzureichend ausgeführt werden, kann es in der Folge zu fehlerhaften Zuständen und Problemen in Wirtschaftszyklen kommen.

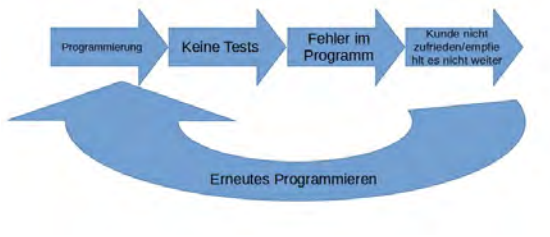


Abb. 1: Das Diagramm zeigt den Kreislauf bei fehlenden Tests [1]

Ein praktisches Beispiel hierfür ist der Vorfall von Crowdstrike, bei dem mehrere Millionen Systeme weltweit betroffen waren und auch kritische Infrastrukturen, wie Krankenhäuser und Flughäfen, zum Erliegen gekommen sind [2]. Vorfälle dieser Art geben Anlass, verschiedene Testmethoden und Testverfahren zu analysieren und zu bewerten.

Testmethoden

Es gibt bereits eine Vielzahl an Testverfahren, die in der aktuellen Zeit verwendet werden:

- Funktionale Tests
- Nicht funktionale Tests
- Statische Tests
- Dynamische Tests

Bei funktionalen Tests wird das Programm einfach auf seinen korrekten Ablauf überprüft. Es wird festgestellt, ob die vorher definierten Anforderungen durch den Ablauf des Programms erwartungsgemäß erfüllt werden. Im Gegensatz dazu sind nicht funktionale Tests darauf ausgelegt Szenarien abzudecken, die auf die Umgebung abzielen. Dies betrifft zum Beispiel den Aufwand an Ressourcen oder die Frage danach, welchen Grad von Intuitivität das User Interface aufweist. Oder Kriterien wie Benutzerfreundlichkeit und User Experience oder Sicherheit, bei denen potentielle Schwachstellen identifiziert werden könnten. Anders als bei den vorherigen Verfahren sind statische Tests auf den Quellcode bezogen. Der Quellcode wird begutachtet und anhand dessen werden bereits Fehler und Schwachstellen erkannt. So lassen sich die Fehler frühzeitig korrigieren, und es kann viel Geld gespart werden. In der Software Entwicklung werden Fehler stets teurer, je später sie erkannt werden. Für dieses Verfahren werden üblicherweise Tools benutzt, die den Programmcode scannen und anschließend Feedback geben. Neben den statischen Tests existieren auch die dynamischen Tests. Bei diesen geht es darum, dass das Programm mit verschiedenen Eingaben gefüttert und ausgeführt wird. Es werden sowohl Szenarien geschaffen, bei denen der Quellcode sichtbar ist als auch Szenarien, bei denen er nicht sichtbar ist. Der Endbenutzer ist üblicherweise ein reiner Anwender und muss den Code hinter der Anwendung nicht verstehen. Deshalb ist es umso wichtiger zu gewährleisten, dass das Programm

von außen gesehen korrekt funktioniert und das zu erwartende Ergebnis auch erzielt wird [4]. Genauere Spezifikationen dazu lassen sich in der ISO 29119 finden, welches ein komplettes Kapitel zu Software Tests beinhaltet [3] .

Spezifische Verfahren

Bei den dynamischen Verfahren gibt es noch weitere Unterscheidungen. Neben Black- und Whitebox-Textverfahren, bei denen der Code entweder nicht vorliegt (Blackbox) oder einsehbar ist (Whitebox), existieren auch die erfahrungsbasierten Tests. Dabei werden Testfälle definiert, die bereits aus vorangegangen Tests resultieren und auch von Personen mit Erfahrung beigesteuert werden. Für Tests können Frameworks benutzt werden, die dafür ausgelegt sind. Viele Programmiersprachen kommen bereits mit vorinstallierten Frameworks, wie JUnit für Java oder auch xUnit für C#. Aber auch End-to-End Tests lassen sich mit Hilfe von Frameworks erstellen. In vielen Fällen werden dafür Selenium und Playwright benutzt. Im Falle einer Website einer Hochschule lassen sich mit ihnen Funktionen im Studentenbereich darauf überprüfen, ob diese die gewünschten Funktionalitäten erfüllen. Dies bezieht sich beispielsweise auf die korrekte Ausgabe von Bescheinigungen und Dokumenten, von Stammdaten sowie deren Validierungen.

Vorgehensweise

Als Teil der Bachelorarbeit sind mit Hilfe der Frameworks Selenium und Playwright Testfälle verfasst worden, die das interne Studentenportal einer Hochschule auf ihre Funktionalitäten testen. Anhand von zuvor erstellen Use Cases sind Szenarien erstellt worden, die einer manuellen Benutzung durch Studenten gleichzusetzen sind. Dabei wird durch mehrfache Ausführungen auch die Fehleranfälligkeit des Systems auf die Probe gestellt. Hierfür wurde der Code der Website intensiv von außen untersucht.

Abb. 2: Bilder zu den Frameworks [1]

Basierend auf diesen Ergebnissen werden die Frameworks anschließend analysiert und dementsprechend gegeneinander beurteilt. Die voraussichtlichen Ergebnisse bestätigen, dass das Studentenportal korrekt und mit vernachlässigbaren Fehlern funktioniert. Darüber hinaus zeigt sich, dass sich - trotz unterschiedlicher Eigenschaften und Strukturen - beide Frameworks hervorragend zur Nutzung von End-To-End Tests nutzen lassen.

Ausblick

In der heutigen Zeit ist es unausweichlich geworden, Testverfahren zur Verifikation von Programmcode einzusetzen. Zwar könnte man durch einen Verzicht auf Tests in der Theorie Zeit einsparen, jedoch würde es früher oder später zu großen Problemen kommen, die bei den beteiligten Firmen und Unternehmen zusätzliche Zeit und immense Kosten verursachen würden. Da menschliche Überprüfungen jedoch extrem zeitaufwändig und zudem fehleranfällig sind, sind automatisierte Testverfahren auf unterschiedlichen Ebenen der richtige Weg zu verbesserter Quantitäts- und Qualitätssicherung. Am Ende erspart es Firmen und Unternehmen zusätzliche Kosten und Zeit, die sie bei fehlerhaften Programmen wieder aufwenden müssen.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Sayonara De Zoysa. Microsoft global outages caused by CrowdStrike software glitch. https://www.researchgate.net/profile/Sayonara-De-Zoysa/publication/382625564_Microsoft_global_outages_caused_by_CrowdStrike_software_glitch/links/66a61b9c4433ad480e80d589/Microsoft-global-outages-caused-by-CrowdStrike-software-glitch.pdf, 07 2024.

[3] International Organization for Standardization. *ISO 29119-4:2021-10, Software- und Systemengineering - Software-Test - Teil 4: Testtechniken. Englischer Titel: Software and system engineering - Software testing - Part 4: Test techniques.* International Organization for Standardization, 2021.

[4] Andreas Spillner and Tilo Linz. *Basiswissen Softwaretest. Aus- und Weiterbildung zum Certified Tester, Foundation Level, nach ISTQB® - Standart.* dpunkt.verlag, 7 edition, 2024.

Ableitung normativer Anforderungen des ”ersetzenden Scannens” unter Berücksichtigung der GoB

Daniel Rupp Fernandes **Catharina Kriegbaum-Kling**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma deron services GmbH, Filderstadt

Einleitung

Prozessdigitalisierung bedeutet, dass bisher analog vorliegende Informationen digital verfügbar gemacht und/oder Arbeitsschritte unter Einsatz von digitaler Technologie ausgeführt werden. Durch das Überführen analoger Informationen in das Digitale, wird die Transparenz erhöht, Arbeitsaufwand und Kosten reduzieren sich, und die Effizienz des Prozesses steigt. [7]

Medienbruch

Ein Medienbruch entsteht, sobald das Übertragungsmedium innerhalb der Prozesskette gewechselt wird, z.B. wenn digitale Dokumente ausgedruckt und wieder eingescannt werden. Ein Medienbruch geht mit dem Risiko des Informationsverlusts und der Informationsverfälschung einher, wie es z.B. bei Tippfehlern innerhalb händisch übertragener Daten der Fall ist. Medienbrüche sind daher beim Konzipieren digitaler Prozesse zu vermeiden bzw. zu eliminieren. In Prozessen der Finanzbuchhaltung treten Medienbrüche besonders häufig auf. Das liegt daran, dass an Dokumente in diesem Bereich besonders hohe rechtliche Anforderungen gestellt werden. Viele Unternehmen setzen deshalb auf Papier, da z.B. Stempel und händische Unterschriften als besonders fälschungssicher gelten und somit einen Vertrauensvorschuss genießen. [6]

Ersetzendes Scannen

„Ersetzendes Scannen“ beschreibt den Vorgang des elektronischen Erfassens von Papierdokumenten (Erstellen eines digitalen Zwillings) und dem anschließenden Vernichten des Originals. Das ersetzende Scannen ist eine gängige Methode, um Medienbrüche in Prozessen abzubauen, da der hierbei entstehende „digitale Zwilling“ anschließend digital, an Stelle des Papiers tritt, und in digitalen Informationssystemen weiterverarbeitet werden kann. Dieser Prozess muss ganzheitlich betrachtet werden, denn die Prozess- und Datenverantwortung beginnt und endet in der Supply-Chain außerhalb des Unternehmens. Das macht die Definition besonders aufwendig, da auch Fehlerbehandlungsroutinen bedacht werden müssen, die in der Prozesskette vor und nach der Unternehmensdomäne liegen. [5]

Normative Anforderungen

Diese ergeben sich aus den Grundsätzen ordnungsgemäßer Buchführung (GoB), welche im Handelsgesetzbuch (HGB) und in der Abgabenordnung (AO) konkretisiert werden. Die GoBD (Grundsätze zur ordnungsmäßigen Führung und Aufbewahrung von Büchern, Aufzeichnungen und Unterlagen in elektronischer Form) erweitert und spezifiziert die GoB für Prozesse, die digitale Informationsverarbeitung einsetzen und ist daher auf das „ersetzende Scannen“ anzuwenden. Aufbewahrungsplicht Nach HGB §238 und §257 gelten für kaufmännische Dokumente Aufbewahrungsfristen, die für Papier- wie auch digitale Dokumente gelten. Die AO regelt in §147 die Art der aufzubewahrenden Dokumente und deren jeweilige Aufbewahrungsfrist, welche in der Regel zwischen 6 und 10 Jahren beträgt. [1]

TR-03138 „RESISCAN“

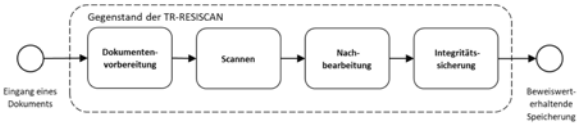


Abb. 1: Der ”generische Scanprozess-[3]

Die Technische Richtlinie „RESISCAN“ (TR-RESISCAN) bietet einen Handlungsleitfaden zur rechtssicheren Gestaltung von Prozessen und technischen Systemen für das ersetzende Scannen. Sie konkretisiert die Anforderungen der GoBD und

empfiehlt Maßnahmen zur praktischen Umsetzung des Scanprozesses. Dies umfasst u.a. die Erstellung einer Verfahrensdokumentation in Form einer detaillierten Beschreibung der Umsetzung technischer, organisatorischer und personeller Maßnahmen entlang des gesamten Prozesses, sodass sich berechtigte Dritte (z.B. Wirtschaftsprüfer) in angemessener Zeit einen ganzheitlichen Überblick verschaffen können. Zusätzlich müssen in Form eines Netzplans alle entlang des Prozesses eingesetzten IT-Systeme, Netze und Anwendungen dargestellt werden. Zuletzt müssen Sicherheitsmaßnahmen konzipiert werden, welche Revisionsschutz (sichere Archivierung über den gesamten gesetzlich vorgeschriebenen Zeitraum) und Qualitätssicherung (die digitalen Dokumente müssen den Anforderungen an Lesbarkeit und Vollständigkeit nach GoB genügen) gewährleisten. [3]

Ziel dieser Arbeit

Ziel dieser Arbeit ist die Einbettung einer Prozessbibliothek des „ersetzenden Scannens“ in die Gesamtprozesskette der Finanzbuchhaltung bei der deron GmbH. Als Handlungsleitfaden für die Umsetzung und Gestaltung rechtssicherer Prozesse und Systeme für das ersetzende Scannen wird die TR-03138 „RESISCAN“ herangezogen. Teil dieser Arbeit ist u.a. die Erstellung der von der TR-03138 geforderten Dokumente, sodass eine Zertifizierung der Prozessbibliothek nach TR-03128 erfolgen kann.

Technische Umsetzung

Der Prozess (siehe Abb. 2) beginnt außerhalb des Unternehmens mit der Erstellung eines für die Buchhaltung relevanten Dokuments, z.B. einer Eingangsrechnung. Mit Eintritt in die Unternehmensdomäne geht die Daten -verantwortung und -Haftung auf den Empfänger über. Daher erfordert der Erhalt des Dokuments geeignete Prüfungs- und ggf. Fehlerbehebungsroutinen, um z.B. mit fehlerhaften Belegen umgehen zu können. Die zu prüfenden Kriterien sind der Verfahrensdokumentation zu entnehmen. Der Beleg kann als digitales Dokument (z.B. PDF) oder in Papierform eingehen. Bereits digitale Dokumente können direkt elektronisch weiterverarbeitet werden. Papierdokumente werden ersetzend gescannt und deren digitaler Zwilling hinsichtlich Qualität und Vollständigkeit geprüft. Nach der Verbuchung im unternehmenseigenen ERP wird das digitale Objekt dauerhaft auf den Servern des Steuer- und Finanz-Dienstleisters DATEV revisionssicher für die vorgeschriebene Zeit gespeichert .

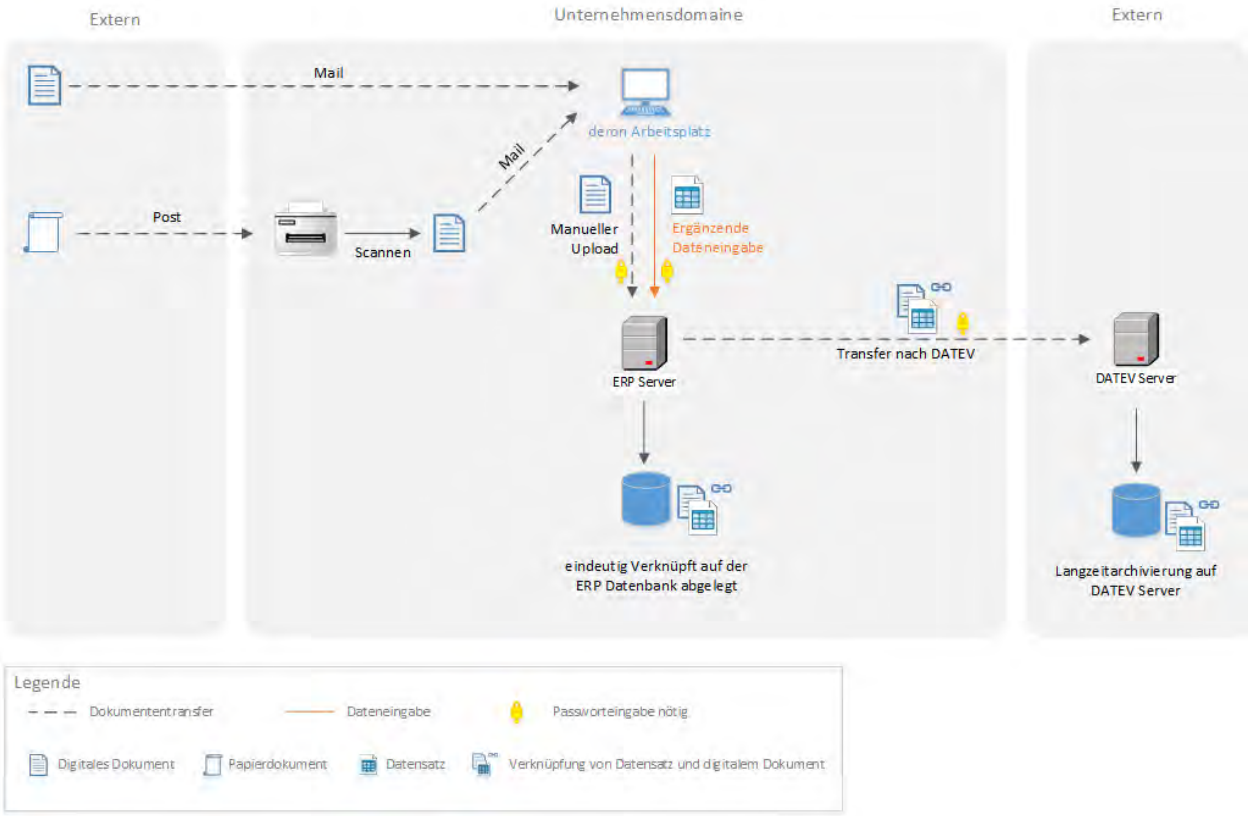


Abb. 2: Darstellung des Informationsfluss entlang des ganzheitlichen Prozesses [2]

Zukunftsausblick

Zwar verspricht die Neuregelung zur obligatorischen elektronischen Rechnung durch das Wachstumschancengesetz ab 2025 einen wesentlichen Baustein zur Digitalisierung des Geschäftsverkehrs. [4] Trotzdem werden uns papiergebundene Verfahren in Deutschland noch lange begleiten, denn ein "Bewirtungsbeleg", ein Parkticket, ein Fahrschein am Automaten, o.Ä. wird noch lange keine Prozessintegration in digitale Pfade der Unternehmen finden. Auch bei der Integration von Rechnungsverfahren aus dem Ausland bleibt das Papier nicht selten der kleinste gemeinsame Nenner.

Literatur und Abbildungen

[1] Gerd Bichler et al. *GoBD Ein Praxisleitfaden für Unternehmen*. AWV e.V:, 2023.

[2] Eigene Darstellung.

[3] Bundesamt für Sicherheit und Informationstechnik. BSI Technische Richtlinie 03138 Ersetzendes Scannen. https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Publikationen/TechnischeRichtlinien/TR03138/TR-03138_V_5.pdf?__blob=publicationFile&v=5, 08 2024.

[4] Haufe-Lexware GmbH und Co KG. Wachstumschancengesetz verkündet. https://www.haufe.de/steuern/gesetzgebung-politik/wachstumschancengesetz_168_600636.html, 03 2024.

[5] Moritz Hübsch. Rechtssicheres ersetzendes Scannen Zur Zulässigkeit von Scanning und Beweiskraft gescannter Dokumente nach Einführung der TR-RESCISCAN. *Computer und Recht*, page 206, 2014.

[6] Leonie Munke. Medienbrüche adé – Wie du Prozesse ganzheitlich digitalisierst. https://www.d-velop.de/blog/prozesse-gestalten/medienbrueche/, 2022.

[7] Florian Reingräber. Prozessdigitalisierung: So bringen Sie Ihr Unternehmen voran. https://scolution.de/prozessdigitalisierung/, 04 2023.

Künstliche Intelligenz in der Psychotherapeutischen Behandlung: Optimierung der Anamnese und Therapieplanung

Mohamad Saleh **Thao Dang**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Mindsetr, Stuttgart

Einleitung

Die mentale Gesundheit gewinnt in unserer Gesellschaft zunehmend an Bedeutung, während Stress und psychische Belastungen weiter zunehmen. Fortschritte in der Technologie, insbesondere im Bereich der Künstlichen Intelligenz (KI) und Large Language Models (LLMs), eröffnen neue Möglichkeiten, psychologische Unterstützung breiter und zugänglicher zu gestalten. Ein zentrales Problem im deutschen Gesundheitssystem ist die Überlastung von Psychotherapeutinnen, die viel Zeit für Erstgespräche aufwenden müssen, um neue Patientinnen einzuschätzen [4]. Die vorliegende Arbeit befasst sich mit der Entwicklung eines KI-gestützten Systems, welches Informationen über den Patienten sammelt und diese in Form eines Anamnesebogens an den Therapeuten weiterleitet.

Ziel der Arbeit

Das Ziel dieser Arbeit ist die Konzeption und Entwicklung eines Prototyps für eine mobile Anwendung, die einen LLM-basierten virtuellen Begleiter in Form eines Chatbots integriert. Dieser virtuelle Begleiter soll Nutzern eine empathische und unterstützende Interaktion bieten, während im Hintergrund relevante Informationen erfasst werden, die eine erste Anamnese ermöglichen. Der resultierende Anamnesebogen wird anschließend den Pschotherapeuten zur Verfügung gestellt, um den Prozess der Erstdiagnose zu automatisieren und ihre Arbeitsbelastung zu reduzieren. Durch die automatisierte Anamnese wird eine präzisere Zuordnung von Patienten zu Psychotherapeut ermöglicht. Dies soll insbesondere die Anzahl erfolgloser Erstgespräche verringern, bei denen Patienten nicht behandelt werden können, weil der Therapeut feststellt, dass der Fall nicht in seinen Fachbereich fällt. Für die Psychotherapeuten bietet das System eine Unterstützung bei der Entscheidungsfindung. Gleichzeitig wird angestrebt, die Wartezeiten für Patienten zu verkürzen und den Zugang zur psychotherapeutischen Versorgung zu verbessern. Ein weiterer Schwerpunkt liegt auf der Skalierbarkeit der Lösung, um sicherzustellen, dass die Anwendung auch in einem Start-up-Kontext erfolgreich umgesetzt und breit eingesetzt werden kann.

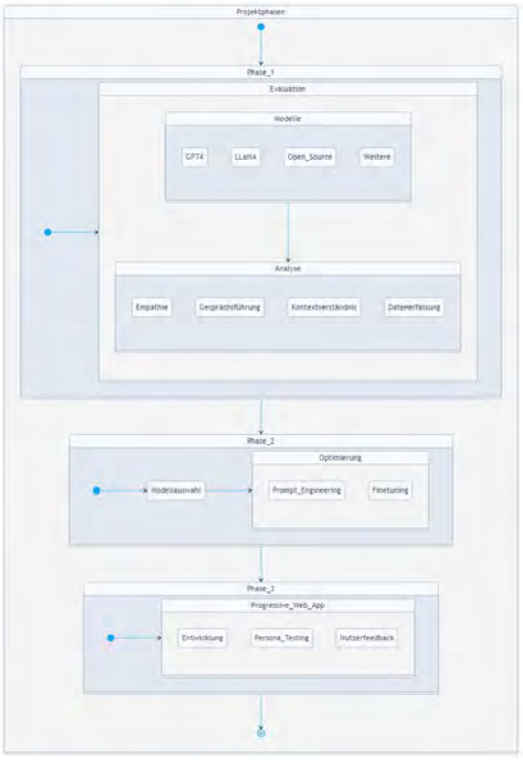


Abb. 1: Mehrstufiger Ansatz zur Realisierung des Projekts [3]

Methodik

Für die Realisierung des Projekts wird ein mehrstufiger Ansatz gewählt, wie in Abbildung 1 dargestellt. In der

ersten Phase werden verschiedene bereits existierende LLMs wie GPT-4, LLaMA und Open-Source-Modelle evaluiert, um deren Fähigkeiten in Bereichen wie Empathie, Gesprächsführung, Kontextverständnis und strukturierter Datenerfassung zu analysieren. Dabei liegt besonderes Augenmerk auf der Qualität der therapeutischen Interaktion und der Fähigkeit, relevante anamnestische Daten zu erfassen. Nach dieser initialen Evaluierung erfolgt eine detaillierte Analyse verschiedener Modelle unter Berücksichtigung von Lizenzierungsoptionen, technischen Anforderungen und Kostenaspekten. Da eine komplette Neuentwicklung eines Modells aus ressourcentechnischen Gründen nicht infrage kommt, wird die Entscheidung auf ein Open-Source-Modell getroffen. Open-Source-Modelle sind bereits vortrainiert, was zu einer Reduktion des Trainingsaufwands führt. Zur Optimierung der Modellleistung werden Prompt-Engineering-Maßnahmen durchgeführt, basierend auf [6]. Diese Techniken tragen dazu bei, die Interaktionen mit LLMs zu strukturieren und die Qualität der generierten Antworten zu steigern, wodurch die therapeutischen Anforderungen an das Modell besser erfüllt werden. Zusätzlich kommen Fine-Tuning-Ansätze zum Einsatz, darunter sowohl Full Finetune als auch die Methode der Low-Rank Adaptation (LoRA). Die letztere Methode erlaubt es, Modelle effizient anzupassen, indem nur ein Bruchteil der ursprünglichen Parameter optimiert wird, was den Speicherbedarf und die Trainingszeit erheblich reduziert. Durch diese Anpassungen wird die therapeutische Kompetenz des Modells gezielt verbessert [5]. Parallel zur Modellauswahl und -optimierung wird ein Prototyp als Progressive Web App (PWA) entwickelt. Diese ermöglicht Plattformunabhängigkeit und wird verschiedenen Nutzern bereitgestellt. Außerdem werden Änderungen direkt an alle Nutzer übertragen, wodurch schnell auf das Feedback der Nutzer eingegangen werden kann. Diese Erkenntnisse fließen dann in die spätere Migration und Implementierung ein [1].

Ergebnisse

Es wurde ein spezielles LLM ausgewählt und durch gezieltes Fine-Tuning für den therapeutischen Kontext optimiert. Dabei wurde Fokus auf einen empathische Kommunikationsmuster gelegt. Der entwickelte PWA-Prototyp demonstriert erfolgreich die geplante Benutzeroberfläche anhand von Mock-Daten. Die Architektur wurde so gestaltet, dass die Integration des optimierten LLMs ermöglicht wird. Zur Evaluation wurden verschiedene Personas entwickelt, die unterschiedliche Patientenprofile und Nutzungsszenarien abbilden. Ein Beispiel eines Personas kann man in Abbildung 2 finden.

Demografische Informationen

Persona Name	Laura Meier
Rolle	Marketing-Managerin
Familienstand	Verheiratet, ein Kind
Alter	32
Einkommen monatlich	3000€ Netto
Bildung	Studium BWL
Wohnort	Berlin

Bibliographie

Laura ist in einer kleinen Stadt aufgewachsen und zog nach Berlin, um an einer renommierten Universität Betriebswirtschaftslehre zu studieren. Nach ihrem Abschluss startete sie ihre Karriere in einer großen Marketingagentur, wo sie schnell aufstieg. Laura ist bekannt für ihre starke Arbeitsmoral und ihre Fähigkeit, mehrere Projekte gleichzeitig zu managen. Doch der ständige Druck und die hohen Erwartungen haben bei ihr zu Stress und Angstzuständen geführt. Laura liebt es, Zeit mit ihrer Familie zu verbringen und sucht nach Wegen, ihre mentale Gesundheit zu verbessern, um sowohl privat als auch beruflich ausgeglichen zu bleiben.

Ziele	Persönliche Probleme
• Laura möchte schnell und einfach Termine bei einem Therapeuten finden	• hoher Stress • Angstzustände
Schmerzpunkte	**Motivationen**
• Lange Wartezeiten auf einen Therapieplatz. • Unsicherheit bei der Suche nach einem passenden Therapeuten.	• Ihre mentale Gesundheit verbessern, um besser mit ihrem stressigen Alltag umgehen zu können. • Effiziente und unkomplizierte Lösungen, die Zeit sparen.

Abb. 2: Beispiel eines Personas [3]

Ausblick

Das LLM wird auf Grundlage der zuvor beschriebenen Personas in den geplanten Nutzertests eingesetzt. Diese Tests haben zum Ziel, sowohl die Qualität der Interaktionen mit dem LLM als auch die Benutzerfreundlichkeit der Anwendung systematisch zu evaluieren [2]. Ein weiterer Fokus zukünftiger Entwicklungen liegt auf der Optimierung des zugrunde liegenden LLMs. Hierbei sollen durch gezieltes Fine-Tuning robustere Antworten sowie eine verbesserte Verarbeitung von Nutzereingaben erzielt werden. Darüber hinaus ist die Erweiterung des Modells um multimodale Fähigkeiten vorgesehen, um Text-, Sprach- und Bilddaten effizient verarbeiten zu können. Auf Basis des Prototyps kann die Implementierung der Software erfolgen. Neue Features sollen die Benutzererfahrung gezielt verbessern und den Funktionsumfang der Anwendung erweitern. Dabei wird dem Aspekt des Datenschutzes besondere Aufmerksamkeit gewidmet, um die Sicherheit und Privatsphäre der Nutzer zu gewährleisten. Langfristig wird eine Migration auf native Apps in Betracht gezogen, um die Stabilität und Performance der Anwendung weiter zu steigern.

Literatur und Abbildungen

[1] Biørn-Hansen Andreas, A. Majchrzak Tim, and Grønli Tor-Morten. Progressive Web Apps for the Unified Development of Mobile Applications., 2018.

[2] Maze co. A Beginner's Guide to Usability Testing. https://maze.co/guides/usability-testing/, 2023.

[3] Eigene Darstellung.

[4] Wissenschaftliche Dienste des Deutschen Bundestages. Wartezeiten auf eine Psychotherapie Studien und Umfragen. https://www.bundestag.de/resource/blob/916578/53724d526490deea69f736b1fda83e76/WD-9-059-22-pdf.pdf, 2022.

[5] J. Hu Edward, Shen Yelong, Wallis Phillip, Allen-Zhu Zeyuan, Li Yuanzhi, Wang Shean, Wang Lu, and Chen Weizhu. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of the International Conference on Learning Representations (ICLR)*. None, 2022.

[6] White Jules, Fu Quchen, Hays Sam, Sandborn Michael, Olea Carlos, Gilbert Henry, Elnashar Ashraf, Smith Jesse Spencer, and Schmidt Douglas C. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT, 2023.

Software der intelligenten Parkraumüberwachung

Nail Sarikaya **Clemens Klöck**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

In der heutigen Zeit sind künstliche Intelligenzen und das maschinelle Lernen aufgrund der zunehmenden Digitalisierung und des schnellen technologischen Fortschritts in verschiedenen Anwendungsbereichen unerlässlich. Einer der wichtigsten Sektoren, in denen KI tätig ist, ist die Bilderkennung, die aufgrund des Einsatzes von **Convolutional Neural Networks (CNN)** sowie anderer neuronaler Netze große Fortschritte gemacht hat. Diese speziellen Netze haben ihre Wirksamkeit bei der automatisierten Verarbeitung und Analyse visueller Daten unter Beweis gestellt und sind in verschiedenen Bereichen, darunter autonomes Fahren, medizinische Diagnostik, Überwachungssysteme und viele andere, weit verbreitet. Das Ziel dieses Projekts ist es, durch den Einsatz von Deep-Learning-Technologien wie **YOLO (You Only Look Once)** und unterstützenden Bibliotheken wie OpenCV und DeepSORT, eine effiziente Objekterkennung und Verfolgung zu entwickeln.

Problemstellung und Zielsetzung

Im Rahmen des Projekts zur intelligenten Parkplatzerkennung besteht die Herausforderung darin, eine Softwarelösung zu entwickeln, die mithilfe von Kameradaten freie und belegte Parkplätze in Echtzeit erkennt und klassifiziert. Dabei müssen die Algorithmen nicht nur zuverlässig, sondern auch ressourcenschonend auf Hardware wie dem Raspberry Pi oder ähnlichen Plattformen laufen. Das Ziel dieser Arbeit ist es, eine Softwarelösung zu entwickeln, die mithilfe von künstlichen neuronalen Netzen (KNN) freie und belegte Parkplätze auf einem Parkplatz erkennen kann. Die Software soll in der Lage sein, Bild- und Videodaten effizient zu verarbeiten, um in Echtzeit Informationen über den Status von Parkplätzen bereitzustellen. Diese Informationen sollen über eine benutzerfreundliche Schnittstelle verfügbar gemacht werden, um die Parkplatzsuche für Endbenutzer zu erleichtern und die Effizienz des Parksystems zu steigern.

Künstliche Neuronale Netzwerke (KNN)

Künstliche Neuronale Netzwerke (KNN) sind mathematische Modelle, die von der Struktur und Funktionsweise des menschlichen Gehirns inspiriert sind. Sie bestehen aus miteinander verbundenen Knoten, den sogenannten Neuronen, die in Schichten organisiert sind: Eingabeschicht, verborgene Schichten und Ausgabeschicht. Diese Netzwerke lernen durch Training, Muster und Zusammenhänge in Daten zu erkennen und komplexe Probleme zu lösen. Anwendungen von KNN reichen von Bilderkennung über Spracherkennung bis hin zur Steuerung autonomer Systeme. [6]

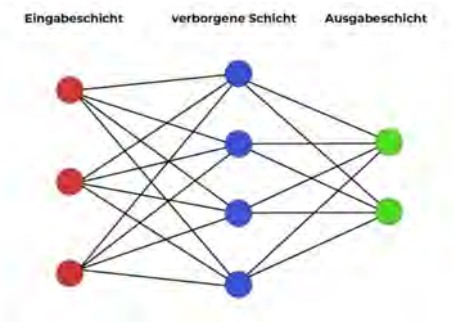


Abb. 1: Aufbau eines Künstlichen Neuronalen Netzes (KNN) [6]

Convolutional Neural Network (CNN)

CNNs spielen eine zentrale Rolle in der Objekterkennung und Autofahrerkennung, insbesondere bei der automatischen Erkennung von Objekten wie Autos, Fußgängern oder Verkehrszeichen in Bildern und Videos. Der Hauptvorteil von CNNs in diesem Kontext liegt in ihrer Fähigkeit, automatisch relevante Merkmale aus Bildern zu extrahieren und hierarchisch zu verarbeiten, um Objekte zuverlässig zu identifizieren. In einem typischen Szenario der Objekterkennung, etwa in einem selbst fahrenden Auto, wird ein Bild oder ein Video-Frame von einer Kamera aufgenommen. Ein CNN analysiert dieses Bild, indem es in mehreren Schichten

durch das Bild „faltet“, um Merkmale wie Kanten, Formen und Texturen zu extrahieren. Diese Merkmale werden in tiefere Schichten weiterverarbeitet, wo das Netzwerk komplexere Muster erkennt, etwa die Form eines Autos oder die Struktur eines Verkehrsschildes. [1] In der Abbildung werden die einzelnen Schichten laut Melanie [3] dargestellt:

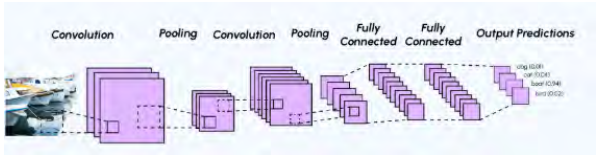


Abb. 2: Convolutional Neural Network Layers [3]

- **Convolutional Layers (Faltungsschichten):** In diesen Schichten wird das Bild mit sogenannten Kernels oder Filtern untersucht, die nach einfachen Mustern wie Kanten, Ecken oder Linien suchen. Diese Filter „gleiten“ über das Bild, um die wichtigen Merkmale zu extrahieren.
- **Pooling Layers (Pooling-Schichten):** Diese Schichten reduzieren die Bildgröße und vereinfachen die Verarbeitung, indem sie nur die markantesten Merkmale aus einem Bereich des Bildes beibehalten (häufig wird Max-Pooling verwendet). Dadurch wird das Netzwerk robuster gegenüber kleinen Verschiebungen und Verzerrungen im Bild.
- **Fully Connected Layers (Vollständig verbundene Schichten):** Am Ende des CNNs werden die extrahierten Merkmale kombiniert, um zu entscheiden, welches Objekt im Bild erkannt wurde. Dies könnte zum Beispiel die Klassifizierung eines Fahrzeugs oder eines Fußgängers sein.

You Only Look Once (YOLO)

YOLO ist ein Algorithmus zur Objekterkennung in Bildern oder Videos. Im Gegensatz zu älteren Methoden, die das Bild in mehrere Teile unterteilen und jedes separat analysieren, betrachtet YOLO das gesamte Bild auf einmal. Dadurch kann es schneller Objekte erkennen, indem es in einem einzigen Durchgang sowohl die Position als auch die Klasse der Objekte vorhersagt. YOLO ist besonders effizient und wird häufig in Echtzeitanwendungen wie Videoüberwachung oder autonomen Fahrzeugen eingesetzt. [4]

DeepSORT mit YOLO

DeepSORT ist ein Algorithmus zur Objektverfolgung, welches YOLO zur Objekterkennung nutzt. Während YOLO Objekte in einem Bild erkennt und Bounding-Boxes liefert, sorgt DeepSORT dafür, dass diese Objekte über mehrere Frames hinweg verfolgt werden. Es verwendet dabei einen Kalman-Filter und visuelle Merkmale, um Objekte eindeutig zu identifizieren und eine stabile Verfolgung auch in komplexen Szenen zu gewährleisten. Die Kombination bietet eine schnelle und zuverlässige Lösung für Echtzeit-Tracking. [5]

Einsatzszenarien

Die Software kann in unterschiedlichen Umgebungen implementiert werden, wie zum Beispiel:

1. **Öffentliche Straßen:** Überwachung von Parkplätzen entlang von Straßen in städtischen Gebieten, um freie und belegte Parkflächen in Echtzeit zu identifizieren.
2. **Wohngebiete:** Erkennung der Parkplatzbelegung in Wohngebieten, um die Verwaltung von Parkflächen zu erleichtern.
3. **Parkhäuser:** Unterstützung von automatisierten Leitsystemen in großen Parkhäusern durch Echtzeitinformationen über belegte und freie Bereiche.

Funktionsweise

Es soll mit einer fest installierten Kamera arbeiten, die Live-Bilddaten in das System einspeist. Die Kombination aus YOLO und DeepSORT ermöglicht es, Fahrzeuge zu erkennen und über mehrere Frames hinweg zu verfolgen. Dabei werden folgende Informationen in Echtzeit verarbeitet:

- Position und Größe von Fahrzeugen.
- Status von definierten Parkbereichen (frei oder belegt).
- Statistiken zur Parkplatznutzung.

Ein einfaches Beispiel, wie es in Parkräumen eingesetzt werden kann, wird in der Abbildung 3 dargestellt.

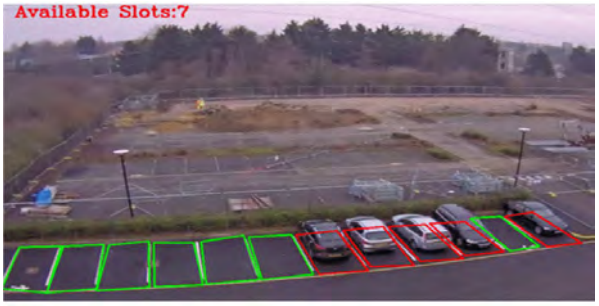


Abb. 3: Parkplatz erkennung mit YOLOv8 [2]

Eigenständige Softwarelösung

Die Software wird so entwickelt, dass sie in Echtzeit Parkplätze erkennen und anzeigen kann, ob diese frei oder belegt sind. Dabei wird darauf geachtet, dass keine zusätzlichen Dienste oder komplizierte Plattformen nötig sind, um die Software nutzen zu können. Ein wichtiger Punkt ist, dass die Daten direkt vor Ort verarbeitet werden sollen, damit die Privatsphäre geschützt bleibt. Ziel ist es, eine einfache, zuverlässige und vielseitige Software zu entwickeln, die in vielen Bereichen eingesetzt werden kann.

Vorteile der Software

- **Echtzeit-Informationen:** Reduzierung der Suchzeit für Parkplätze durch aktuelle Daten.
- **Skalierbarkeit:** Flexible Anpassung an verschiedene Größen von Parkbereichen und Verkehrssystemen.
- **Nachhaltigkeit:** Verringerung des Suchverkehrs und damit des CO2-Ausstoßes in urbanen Gebieten.
- **Automatisierung:** Weniger manueller Aufwand für die Verwaltung von Parkplätzen.

Ausblick

Die Zukunft der Parkplatzerkennung bietet viel Potenzial, insbesondere durch den Einsatz von Deep Learning und Convolutional Neural Networks (CNNs). Mit Fortschritten in der Bildverarbeitung und Echtzeit-Datenanalyse könnte die Technologie zunehmend präziser und effizienter werden, um verfügbare Parkplätze in verschiedenen Umgebungen zuverlässig zu erkennen.

Literatur und Abbildungen

[1] Saad Albawi, Mohammed Tareq Abed, and Saad Al-Zawi. Understanding of a convolutional neural network. https://ieeexplore.ieee.org/document/8308186, 03 2018.

[2] Nagamani Gonthina, Santhosh Katkam, et al. Parking Slot Detection Using Yolov8. https://ieeexplore.ieee.org/document/10435799, 02 2024.

[3] W Melanie. Convolutional Neural Network: Everything You Need to Know. https://datascientest.com/en/convolutional-neural-network-everything-you-need-to-know, 09 2023.

[4] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. https://ieeexplore.ieee.org/document/7780460, 12 2016.

[5] QiFeng Sui. Multi-Target Tracking Based on YOLOv8 and DeepSORT. https://ieeexplore.ieee.org/document/10692499, 10 2024.

[6] Laurenz Wuttke. Künstliche Neuronale Netzwerke: Definition, Einführung, Arten und Funktion. https://datasolut.com/neuronale-netzwerke-einfuehrung/, 02 2024.

Implementierung eines Remote-Zugriffs zur Unterstützung von Testsystemen auf Basis der STM32H7-Plattform

Uwe Schall **Clemens Klöck**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Steinbeis Embedded Systems Technologies GmbH, Esslingen am Neckar

Einleitung

Durch die zunehmende Digitalisierung und der dadurch folgenden Zunahme der Anforderungen werden elektronische Systeme immer komplexer. Diese Steigerung hat zufolge, dass Komponenten und Systeme immer umfangreicher werden. Wo vor Jahren häufig noch einfache Platinenschaltungen eingesetzt wurden, werden heutzutage fast ausschließlich komplexe mehrlagige Platinen mit vielen verschiedenen Komponenten verwendet. Dies hat auch zur Folge, dass Softwarekomponenten immer umfangreicher werden, um der komplexen Hardware und den gewachsenen Anforderungen gerecht zu werden. Um die Qualität und Zuverlässigkeit solcher Systeme zu gewährleisten, gewinnt das Testen während des gesamten Produktlebenszyklus zunehmend an Bedeutung. [1]
Das Testen muss daher im gesamten Produktlebenszyklus, beginnend bei der Entwicklung, bei der Produktqualifizierung, bei der Serienprüfung im Rahmen der Fertigung und bei der Prüfung von Rückläufern, verankert sein. Deshalb ist die Nachfrage nach Testsystemen, die der neu gewonnenen Komplexität gewachsen und trotzdem einfach handhabbar sind, hoch. [4]

Hinführung und Zielsetzung

Im Rahmen der Thesis wird eine vorhandene Testsystem-Software auf einer neuen Hardware-Plattform mit STM32H7-Mikrocontroller implementiert. Dieses Testsystem ermöglicht es, über einen Remote-Zugriff verschiedene Tests auf Mikrocontroller-Systemen durchzuführen. Um das Testen möglichst einfach zu gestalten, abstrahiert dieses vorhandene System Hardwarekomponenten durch die Zuweisung von Labels. So kann etwa eine LED einfach mit LED_GRUEN angesprochen werden, ohne Kenntnis über die genaue Verschaltung der LED auf der Platine zu benötigen. Diese Abstrahierung macht sich Dateien und Dateisysteme zunutze. Für jede zu abstrahierende Hardware wird eine Datei angelegt. In dieser wird das Mapping für die Abstrahierung gespeichert. Beim Start der Applikation werden nun alle Dateien, auch Deskriptoren genannt, geladen. Wird eine Hardware mit einem Label angesprochen, so kann die zuvor geladene erforderliche Einstellung angewandt und die Hardware passend angesprochen werden. Zusätzlich soll für die neue Hardware-Plattform der Remote-Zugriff neben dem seriellen Zugriff über USB auch über das weitverbreitete Kommunikationsmedium Ethernet genutzt werden und somit das Testsystem um einen neuen Kommunikationstyp erweitern.

Architektur und Implementierung des Remote-Zugriffs

Ein Remote-Zugriff ermöglicht die Steuerung und Überwachung eines Systems aus der Ferne. Bei eingebetteten Systemen besteht die größte Herausforderung darin, einen Remote-Zugriff mit den oft begrenzten Ressourcen in akzeptabler Geschwindigkeit umzusetzen. [3]
In der Implementierung wird ein Timer verwendet, der in regelmäßigen Intervallen prüft, ob Aktionen erforderlich sind. Eine Aktion kann zum Beispiel gesetzt werden, wenn Daten empfangen und diese verarbeitet werden müssen. Diese Architektur gewährleistet, dass der Remote-Zugriff ressourceneffizient bleibt und parallel andere Aufgaben durchgeführt werden können. Dieser Aufbau ist in Abbildung 1 zu sehen und beschreibt in den Schritten 1), 2), 3) und 4) den Standardablauf der Implementierung.
Um den Mikrocontroller über den Remote-Zugriff ansprechen zu können, wird auf der PC-Seite Python verwendet. Python ermöglicht durch seine vielseitigen Bibliotheken und einfache Syntax eine schnelle Entwicklung von Tools, um Tests auszuführen, Daten zu visualisieren und Ergebnisse zu analysieren. In einer Python Bibliothek werden alle Funktionalitäten des Remote-Zugriffs zur Verfügung gestellt so dass sie einfach vom Benutzer aufgerufen werden können.

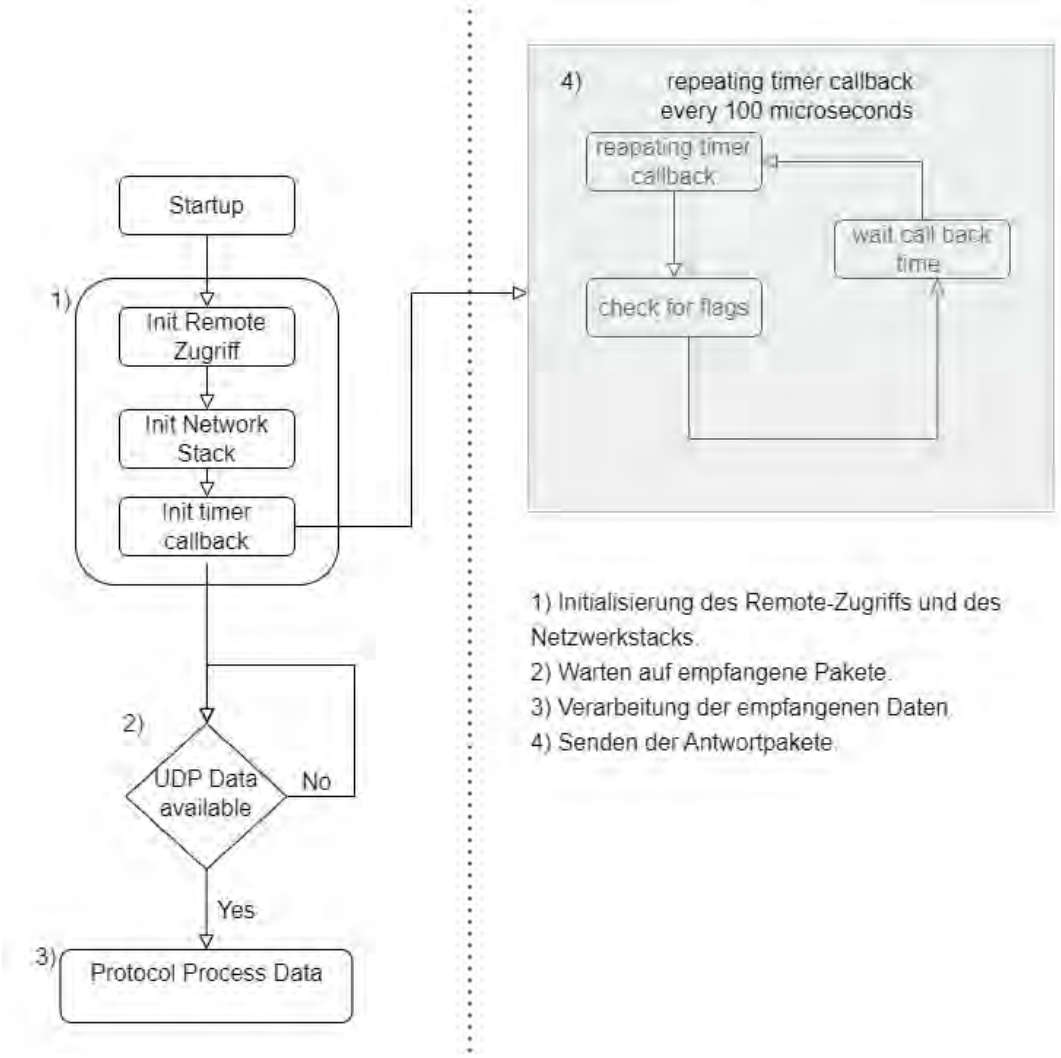


Abb. 1: Programm Ablauf des Remote-Zugriffs [2]

Um im Testkontext wirksam zu sein, bietet der Remote-Zugriff als Hauptfunktion einen Remote Procedure Call an. Dieser ermöglicht es, auf der Hardware eine Funktion mit Übergabeparametern aufzurufen. So können alle zum Testen benötigten Funktionen auf der Hardware zur Verfügung gestellt und dann von außen aufgerufen werden. [5]

Implementierung des Treibers

Um die vorher beschriebene Testplattform auf einem neuen Gerät verwenden zu können, müssen Treiber implementiert werden. Diese stellen geräteabhängige Funktionen zur Verfügung, um Hardwarekomponenten anzusprechen. In dieser Implementierung muss ein Timer zur Verfügung gestellt werden. Außerdem müssen Funktionen zum Senden und Empfangen von Ethernet-Paketen bereitgestellt werden. Um Dateisysteme anlegen und beschreiben zu können, müssen zusätzlich Funktionen zum Schreiben in Flash- und Arbeitsspeicher zur Verfügung gestellt werden.

Benchmarking

Durchgeführte Benchmarks zeigen, dass das System zuverlässig funktioniert und vergleichsweise schnelle Antwortzeiten möglich sind. Als Benchmark wurde ein einfacher Remote Procedure Call ohne Übergabeparameter aufgerufen. Das Timer-Intervall wurde so implementiert, dass es alle hundert Mikrosekunden aufgerufen wird. Das gesamte Paket hat eine Größe von circa sechzig Byte, in diesem Paket sind fünfzehn Byte Nutzdaten enthalten. Die durchschnittliche Antwortzeit beträgt etwa 0.6 Millisekunden, siehe auch Abbildung 2. Dieses Ergebnis ist für den Einsatz im Test von System zufriedenstellend.

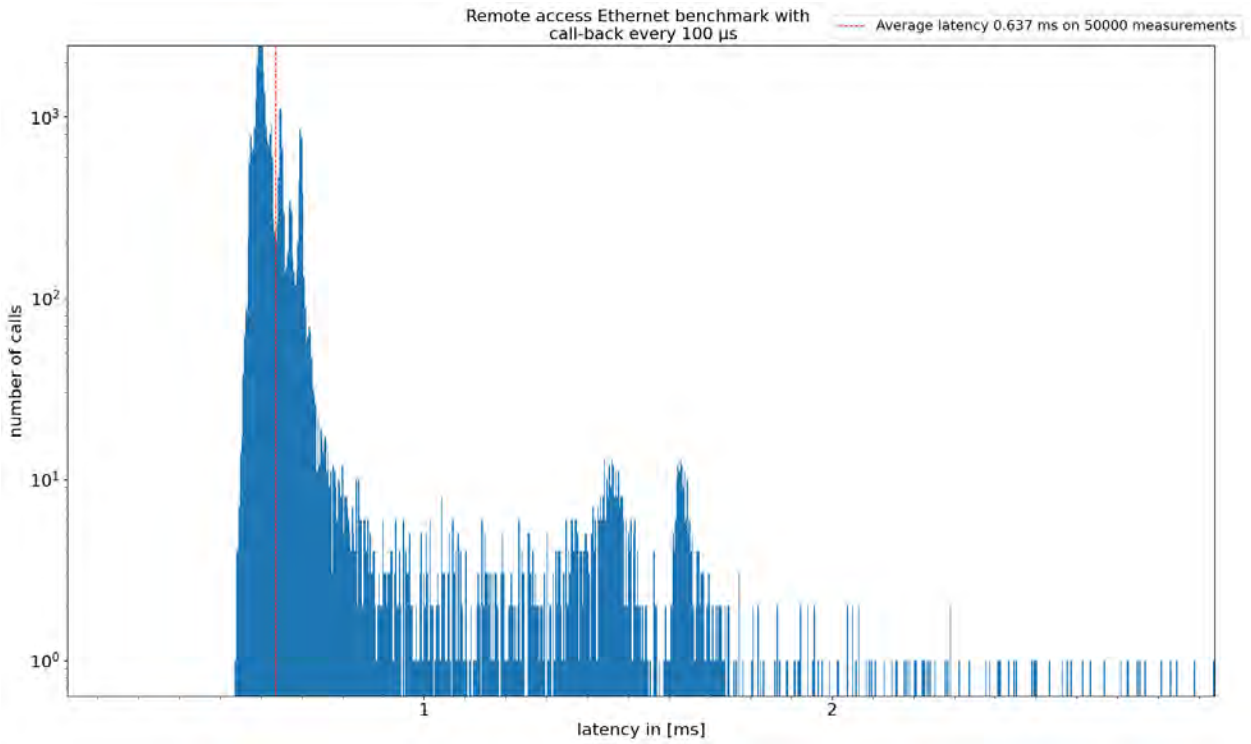


Abb. 2: Benchmark des Remote-Zugriffs [2]

Ergebnis und Ausblick

Durch die Implementierung des Remote-Zugriffs können komplexe Platinenschaltungen einfach und schnell getestet werden. Durch die Abstrahierung komplexer Konfigurationen können die Tests auch ohne großes Wissen über die Hardware durchgeführt werden. Des Weiteren lässt sich das System auch einfach um weitere Technologien, wie das in der Industrie eingesetzte Single-Pair-Ethernet, erweitern. Abschließend kann gesagt werden, dass das System in Zukunft zur zeiteffizienten Testung von Systemen mit Ethernet-Konnektivität eingesetzt werden kann.

Literatur und Abbildungen

[1] E. A. Boldyreva, A. V. Penskoi, and A. E. Platunov. Formalization and Evolution of Learning Path in Embedded Systems. In *2020 Wave Electronics and its Application in Information and Telecommunication Systems (WECONF)*. Engineers, Institute of Electrical and Electronics, 2020.

[2] Eigene Darstellung.

[3] Edward Ashford Lee and Sanjit Arunkumar Seshia. *Introduction to Embedded Systems: A Cyber-Physical Systems Approach*. Massachusetts Institute of Technology Press, 2 edition, 2017.

[4] Peter Marwedel. *Eingebettete Systeme: Grundlagen Eingebetteter Systeme in Cyber-Physikalischen Systemen*. Springer Vieweg, 2 edition, 2021.

[5] Christoph Meinel and Harald Sack. *Internetworking: Technische Grundlagen und Anwendungen*. Springer Vieweg, 2012.

Archivieren von historischen Zeitseriendaten

Timo Schaller **Harald Melcher**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma AW Connectivity Platform GmbH, Nürtingen

Motivation

Im industriellen Internet of Things (IIoT) ist die Erfassung und Speicherung von Prozessdaten von zentraler Bedeutung. Betreiber benötigen sowohl aktuelle als auch historische Einblicke in das Verhalten ihrer Anlagen, um Entscheidungen beispielsweise bezüglich der Wartung treffen zu können. Diese Daten bilden die Grundlage für weiterführende Dienste. Dazu gehören Visualisierungen von Anlagezuständen und das Monitoring. Letzteres umfasst den Prozess der Erfassung, Verarbeitung und Anzeige von Echtzeitdaten. Außerdem ist Alerting ein wichtiger Bestandteil, das über beispielsweise unerwartete Änderungen im Systemzustand informiert und vorausschauende Wartung unterstützt. Die Anforderungen an die Anzahl der aus den Daten resultierenden Datenpunkte und deren Abtastrate unterscheiden sich je nach Prozess stark. Weniger dynamische Prozesse kommen häufig mit Datenerfassungen in Intervallen von 5 oder 15 Minuten aus, während schnellere Prozesse eine Erfassung auf Sekundenbasis erfordern. Bei mehreren Tausend Datenpunkten pro Prozess entstehen rasch umfangreiche Datenmengen, die gespeichert und abgerufen werden müssen. Um diesen Herausforderungen gerecht zu werden, wurden spezielle Zeitseriendatenbanken entwickelt. Im Vergleich zu herkömmlichen relationalen Datenbanken bieten sie Vorteile in Bezug auf Speicherverbrauch, Abfragegeschwindigkeit und Datenaggregation. Bei letzterem handelt es sich um eine Methode, die mehrere Datensätze in einem Abrechnungsdatensatz zusammenfasst.

Problemstellung und Zielsetzung

Trotz dieser genannten Vorteile von Zeitseriendatenbanken entstehen durch die Speicherung großer Datenmengen hohe Kosten. Leistungsstarke Speicherlösungen wie Solid-State-Drives (SSDs) und eine hohe Rechenleistung sind für die Datenverwaltung notwendig. Die Kosten für diese Speicherung stehen oft in keinem Verhältnis zur tatsächlichen Nutzung, und regelmäßige vollständige Backups belasten die Systeme zusätzlich. Aufgrund dieser Gegebenheiten zielt diese Arbeit darauf ab, eine Lösung zu evaluieren, die es ermöglicht, Daten effizient aus einer Zeitseriendatenbank in ein Archiv zu überführen. Dieses Archiv soll den Speicherbedarf sowie die damit verbundenen Kosten reduzieren. Gleichzeitig ermöglicht es eine schnelle Abfrage archivierter Daten, ohne größere Wartezeiten und Serverlast.

Zeitseriendatenbanken

Im Gegensatz zu relationalen Datenbanken, die sich auf strukturierte Daten und festgelegte Beziehungen konzentrieren, sowie zu NoSQL-Datenbanken, die unstrukturierte und semi-strukturierte Daten unterstützen, gewinnen Zeitseriendatenbanken (Time-Series Databases, TSDBs) immer mehr an Bedeutung. Sie sind darauf ausgelegt, große Mengen an zeitgestempelten Daten [6] zu speichern und abzurufen. Das sind Daten, die mit einem bestimmten Zeitpunkt (Timestamp) - bestehend aus Datum und Uhrzeit - versehen sind, zu dem sie erfasst oder erstellt wurden. Durch die zunehmende Digitalisierung, z. B. im Gesundheitswesen und in der Entwicklung intelligenter Städte, wird die Quantität, Qualität und Bedeutung von Zeitseriendaten in den kommenden Jahren rapide ansteigen, argumentiert Nielsen in [7]. Während die Idee, Zeitserien zu erfassen und zu analysieren, nicht neu ist, stellt die enorme Größe moderner Datensätze und die Vielzahl neuer Datenquellen die gegenwärtige Herausforderung dar, skalierbare Zeitseriendatenbanken einzusetzen.

Eigenschaften von Zeitseriendatenbanken

TSDBs bieten entscheidende Vorteile gegenüber traditionellen Datenbanken. Besonders im Umgang mit der enormen Menge an Datenpunkten, die in kurzen Intervallen gesammelt werden. Ein charakteristisches Merkmal dieser Datenbanken ist die Fähigkeit, Daten automatisch zu komprimieren und zu aggregieren, wodurch die Speicherung von Rohdaten verwaltet wird,

ohne an Genauigkeit zu verlieren. Die Möglichkeit, mehrere Beobachtungen über die Zeit hinweg zu speichern und zu analysieren, ist von großem Wert. Die Optimierung auf schnelle Schreibvorgänge ermöglicht es TSDBs, große Datenmengen zu verarbeiten, ohne die Abfrageleistung zu beeinträchtigen. Ein weiteres Element von TSDBs ist auch die Aggregation von Daten über Zeiträume hinweg, was für die Analyse langfristiger Trends unerlässlich ist [4].

InfluxDB als Quelle von Zeitreihendaten

Die von der Firma InfluxData entwickelte Zeitseriendatenbank wurde speziell für die Speicherung und Analyse von Zeitreihendaten konzipiert. Durch ihre Optimierung für die Verarbeitung großer Datenströme mit minimaler Latenz eignet sie sich für Echtzeitanalysen [2]. InfluxDB unterstützt unter anderem die Integration mit offenen Standards wie Apache Parquet. Zudem verwendet InfluxDB Apache Arrow [3], ein plattformübergreifendes In-Memory-Datenformat, das schnelle und effiziente Datenverarbeitung gewährleistet. Auf Basis von Apache Arrow nutzt InfluxDB zudem FlightSQL [5], ein Protokoll zur Datenübertragung zwischen Datenbank und Clients. Die TSDB unterstützt sowohl SQL als auch die SQL-ähnliche Abfragesprache InfluxQL, die speziell für die Interaktion mit InfluxDB entwickelt wurde. Sie ermöglicht es, Zeitseriendaten abzufragen, zu aggregieren und zu manipulieren. Ein typisches Szenario könnte sein, dass man Metriken eines Servers aus einer Influx-Datenbank abfragen möchte. Zum Beispiel könnte eine Abfrage die durchschnittliche CPU-Auslastung (in Prozent) über einen bestimmten Zeitraum, wie die letzten 24 Stunden, betreffen. Dies könnte sich beispielsweise auf die letzten 24 Stunden beziehen.

Evaluation - Leistungsvergleich der Datenformate

Für die Auswahl des geeigneten Datenformats wurden mehrere Formate auf Funktionalität untersucht dazu gehörten Kriterien wie *Speicherplatzbedarf*, *Komplexität*, *Verfügbarkeit von Bibliotheken* und *Verarbeitungsgeschwindigkeit*. Um die Auswahl des Datenformats jedoch nicht nur auf theoretische Untersuchungen zu stützen, sondern auch auf praktische Erfahrungen, wurde die Leistungsfähigkeit der Formate durch eigene Implementierungen untersucht. Daher werden die Auswirkungen der verschiedenen Datenformate auf die Dateigröße und die Extraktionsdauer von 1 Millionen Datensätzen aus einer InfluxDB demonstriert. Hierfür wurden zwei Konsolenanwendungen entwickelt: *InfluxDBDataWriter* und *InfluxDBDataExtractor*. Die InfluxDBDataWriter-Anwendung ermöglicht das Schreiben der 1 Millionen Mock-Datensätzen in InfluxDB. Hierbei können Benutzer zwischen zwei Datensatzformaten wählen:

1. **UseCase 1**: Datapoints mit jeweils 30 Feldern (bestehend aus 10 Bits, 10 Floats und 10 Longs): Diese Datensätze simulieren ein umfassenderes Szenario. Bei den Bits sind 50 % der Werte auf 0 und 50 % auf 1 gesetzt. Alle Float-Werte betragen 2.00, während alle Long-Werte konstant den Wert 10 annehmen. Der Fokus liegt darauf zu untersuchen, ob die Redundanz der Werte Einfluss auf die Komprimierung hat.
2. **UseCase 2**: Datapoints mit jeweils 3 Feldern (Bit, Float und Long): Simulierte Beispieldaten einer Ladesäule für Fahrzeuge. Die Datensätze enthalten zahlreiche Nullsequenzen (0 - 0,00 - 0), die auftreten, wenn die Ladesäule nicht in Benutzung ist. Beim Ladevorgang werden jedoch Werte generiert, die in einem nahen Spektrum zueinander liegen, um realistische Szenarien abzubilden.

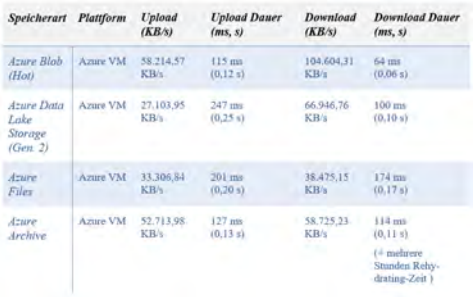
Die InfluxDBDataExtractor-Anwendung extrahiert die vom Writer geschriebenen Datensätze in verschiedenen Formaten und überprüft die Speichergröße sowie die Extraktionsdauer. Die unterstützten Formate umfassen:

- CSV,
- CSV (komprimiert mit gzip),
- JSON,
- JSON (komprimiert mit gzip),
- Apache Avro,
- Apache Parquet und
- Apache ORC

Für die Leistungsanalyse der betrachteten Datenformate wurde die Größe der InfluxDB-Daten mit den entsprechenden komprimierten Dateiformaten verglichen. Der für die Untersuchung zur Grundlage genommene *autogen*-Ordner wird für die Ermittlung der Datenbankgröße verwendet, weil er die automatisch generierten Metadaten und Zeitstempel in InfluxDB enthält, die für die Speicherung von Zeitreihendaten erforderlich sind. InfluxDB speichert Daten in sogenannten *Shards*, die nach Zeitstempel und anderen Kriterien aufgeteilt werden. Der autogen-Ordner ist die Standardzeitreihen-Datenbank, die die wichtigsten und meistgenutzten Daten enthält, die automatisch erstellt werden, wenn keine spezifische Datenbank oder Retention Policy konfiguriert wurde. Die Ergebnisse können der Abbildung 1 entnommen werden, wobei die Entscheidung auf gezippte CSV-Dateien fiel:

Datenformat	Platzbedarf	Komplexität	Verfügbarkeit von Bibliotheken	Verarbeitungs-geschwindigkeit
CSV, **zipped** (textbasiert)				
Apache Parquet (spaltenbasiert)				
Apache Avro (binär)				
JSON, **zipped** (textbasiert)				
CSV (textbasiert)				
JSON (textbasiert)				
Apache ORC (spaltenbasiert)				

Abb. 1: Ergebnisse - Auswahl des Datenformats [1]



Speicherart	*Plattform*	*Upload (KB/s)*	*Upload Dauer (ms, s)*	*Download (KB/s)*	*Download Dauer (ms, s)*
Azure Blob (Hot)	Azure VM	58.214,57 KB/s	115 ms (0,12 s)	104.604,31 KB/s	64 ms (0,06 s)
Azure Data Lake Storage (Gen. 2)	Azure VM	27.103,95 KB/s	247 ms (0,25 s)	66.946,76 KB/s	100 ms (0,10 s)
Azure Files	Azure VM	33.306,84 KB/s	201 ms (0,20 s)	38.475,15 KB/s	174 ms (0,17 s)
Azure Archive	Azure VM	52.713,98 KB/s	127 ms (0,13 s)	58.725,23 KB/s	114 ms (0,11 s) (+ mehrere Stunden Rehydrating-Zeit)

Abb. 2: Vergleich der Upload- und Downloadgeschwindigkeit für Zips [1]

Microsoft Azure als Cloudspeicherlösung

Die Cloud-Plattform *Microsoft Azure* verfügt über eine Vielzahl von Dienstleistungen und Werkzeugen, die für die Verarbeitung, Analyse und Speicherung von Daten erforderlich sind. Da die Nachfrage nach effizienten Datenlösungen ständig wächst, spielt Azure eine wichtige Rolle im IIoT-Bereich. Die Plattform bietet skalierbare Infrastrukturen, die auf die Anforderungen von Big Data und der Datenintegration zugeschnitten sind. Azure ermöglicht die Integration von Prozessdaten durch Echtzeitanalysen, Monitoring und vorausschauende Wartung. Für die Aufgabe, eine geeignete Speichertechnologie zum Archivieren der komprimierten InfluxDB-Daten zu finden, wurden die vier nachfolgenden Speicherarten auf Kosten und Leistungsfähigkeit untersucht und bewertet, wobei der Abbildung 2 die Ergebnisse der Upload- und Downloadgeschwindigkeiten von komprimierten Archivdaten über eine Azure-VM eingesehen werden können:

- Azure Blob Storage (Hot-, Cool-, Cold-Tier),
- Azure Archive Storage (Archive-Tier),
- Azure File Storage und
- Azure Data Lake Storage (Gen2)

Aktueller Stand

Die Evaluation der Speichertechnologien von Microsoft Azure ergab, dass sich Azure Data Lake Storage am besten bzgl. Performance und Kosten für die Archivierung der Zeitseriendaten eignet. Daher wurde nach dieser entscheidung die Umsetzung der Anwendung begonnen. Die hierzu angefertigen .NET-Projekte erfüllen jeweils spezifische Aufgaben erfüllen:

1. **Daten-Upload**: Die *.NET-Konsolenanwendung* **ArchiveUploader** verwaltet den kontinuierlichen Daten-Upload. Sie extrahiert dabei Daten aus der InfluxDB, komprimierte diese und archiviert sie dann im Azure Data Lake Gen2. Der Zugriff auf die InfluxDB erfolgt hierbei einem festlegbaren Intervall, der zeitgesteuert über ei-nen Scheduler getriggert wird. Die Wahl zu einer Konsolenanwendung begründet sich durch eine automatisierte Datenverarbeitung im Hintergrund, ohne Benutzer-eingriff.
2. **Datenabfrage**: Das zweite Projekt **TSDataManager** ist eine *ASP.NET Core Web API*, die für die Abfrage der komprimierten Daten aus dem Azure Data Lake entwickelt wurde. Über den Endpunkt *GET /api/azure_csv_data* können die archivierten Daten abgerufen werden. Für die Visualisierung der Daten wurde eine React-Application implementiert.

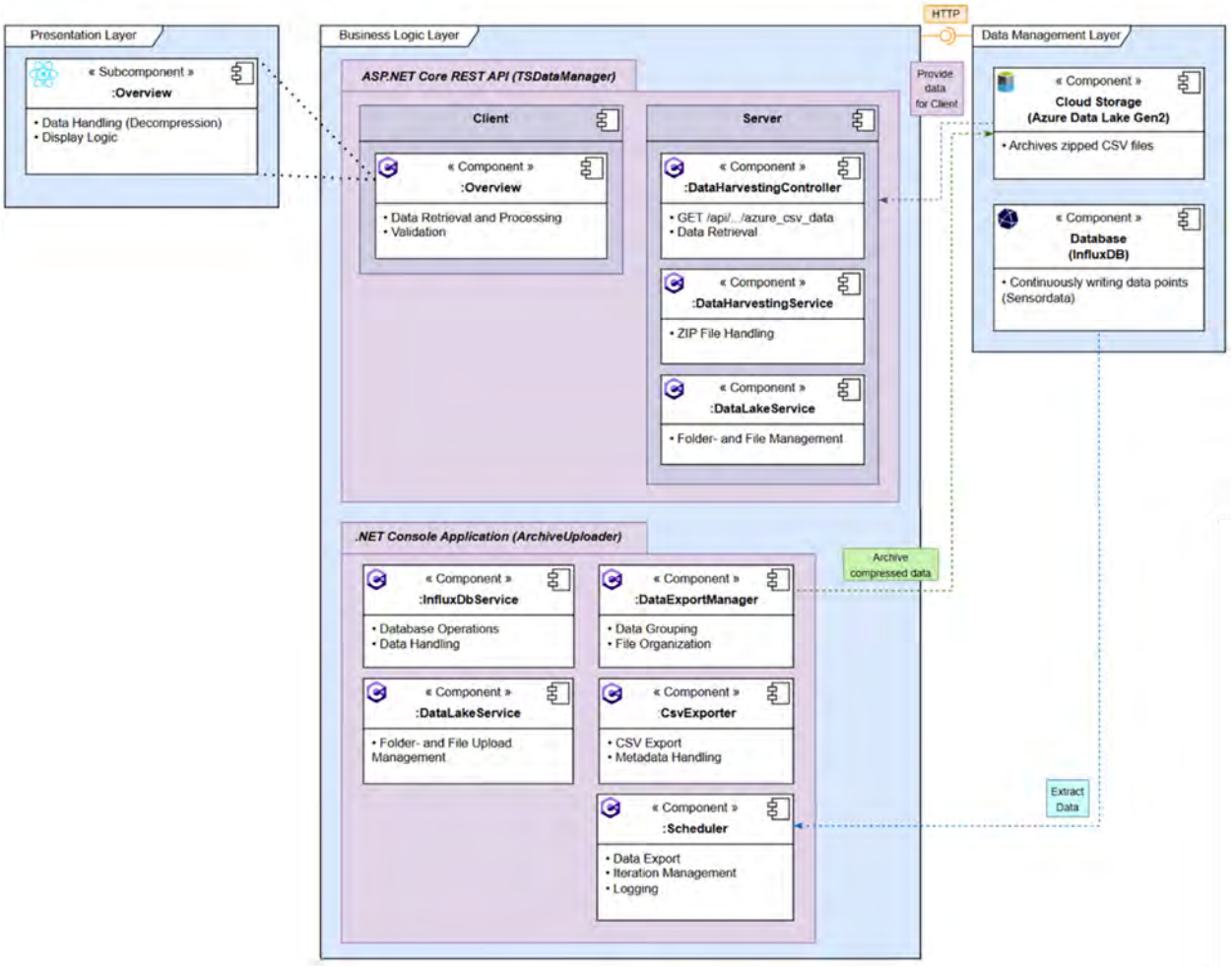


Abb. 3: Architektur (UML-Komponentendiagramm) [1]

Ausblick

Der entwickelte Prototyp und die zuvor durchgeführte Evaluierung haben wesentliche Erkenntnisse über die Archivierung und Verarbeitung von Zeitseriendaten geliefert. Insbesondere der Vergleich verschiedener Datenformate und Speichertechnologien hat gezeigt, wie entscheidend die Wahl der richtigen Lösung für die Effizienz von Daten-managementprozessen ist. Das Projekt bietet somit eine fundierte Grundlage für die künftige Weiterentwicklung und Optimierung von Systemen, die große Datenmengen von Zeitseriendaten verwalten müssen. Trotz bereits identifizierter Optimierungsmöglichkeiten bleibt es eine Herausforderung, diese Technologien in skalierbare und hoch performante Systeme zu integrieren. Zukünftige Entwicklungen könnten verstärkt auf die weitere Verbesserung der Datenkompression und Abfragegeschwindigkeit abzielen, um noch leistungsfähigere Lösungen zu schaffen. Weiterhin könnte die Integration zusätzlicher Technologien und Datenbanken den Prototypen flexibler und anpassungsfähiger machen. Insgesamt hat das Projekt einen wichtigen Beitrag zur effektiven Archivierung von Zeitseriendaten geleistet. Es stellt sowohl einen praktischen als auch einen theoretischen Beitrag dar, der zukünftige Forschung und Entwicklungen in diesem Bereich unterstützen wird.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Paul Dix. Why Time Series Matters For Metrics, Real-Time Analytics And Sensor Data. *An InfluxData Technical Paper*, page 9, 2023.

[3] Anais Dotis-Georgiou. Introduction to Apache Arrow. https://www.influxdata.com/glossary/apache-arrow/, 2023.

[4] Ted Dunning and Ellen Friedman. *Time Series Databases - New Ways to Store and Access Data*. O'Reilly Media, Inc., 2015.

[5] Team InfluxData. Arrow Flight SQL | InfluxData Documentation. https://docs.influxdata.com/influxdb/cloud-dedicated/reference/internals/arrow-flightsql/, 2024.

[6] Team InfluxData. InfluxDB data elements: Timestamp | Documentation. https://docs.influxdata.com/influxdb/cloud/reference/key-concepts/data-elements/, 2024.

[7] Aileen Nielsen. *Practical Time Series Analysis Prediction with Statistics & Machine Learning*. O'Reilly Media, Inc., 2020.

Accelerating AUTOSAR Adaptive Startup with Database-Driven Configuration Data Management

Leon Schmidt **Mirko Sonntag**

Department of Computer Science and Engineering, Esslingen University

Work carried out at Vector Informatik GmbH, Stuttgart

Introduction

AUTOSAR Adaptive is a sophisticated software framework designed to address the increasing demands of advanced automotive applications, such as autonomous driving, connected vehicles, and complex vehicle functions. As an implementation of the AUTOSAR Adaptive Platform, MICROSAR Adaptive by Vector Informatik provides a modular solution with components that are flexible. [1] As MICROSAR Adaptive evolves, performance requirements, particularly regarding startup times, have become a critical area of focus for Original Equipment Manufacturers (OEMs).

Problem Statement

This work explores the feasibility of leveraging database technologies to improve the performance of MICROSAR Adaptive's configuration data management, identifying the limitations of the current JSON-based system and the potential advantages offered by database solutions.

Database vs JSON

MICROSAR Adaptive relies on JSON files for runtime configuration, with each application component using its own file. While JSON offers simplicity and flexibility, it has significant performance drawbacks. An initial analysis reveals two major bottlenecks.
JSON parsing is computationally intensive. A typical application consists of five components that need configuration, parsing the JSON configuration files for all components requires several hundred milliseconds. For a deployment involving 70 applications, this cumulative parsing time extends to several seconds. Reducing this overhead could, in some cases, improve the startup time by up to 50%. Additionally, the current system requires opening multiple JSON files for each application. As the read operations are heavily dependant on the I/O device used, this further impacts the startup latency.

These limitations suggest that the JSON-based configuration approach, while functional, is suboptimal for large-scale deployments with tight startup time requirements.
Database technologies offer a compelling alternative for configuration data management. Widely used in software systems to optimize performance, databases like the Windows Registry exemplify how performance-optimized key-value stores can efficiently manage configuration data.

- Efficient Parsing: Databases store data in a structured format, eliminating the need for repetitive parsing operations. Indexing and optimized query execution further reduce access times.
- Unified Storage: Instead of handling multiple files, a database can centralize configuration data, reducing file I/O overhead.
- Scalability: Databases are designed to handle large volumes of data efficiently, making them suitable for growing application ecosystems.

Design Concepts

Integrating database technologies into the adaptive applications leads to multiple design concepts.

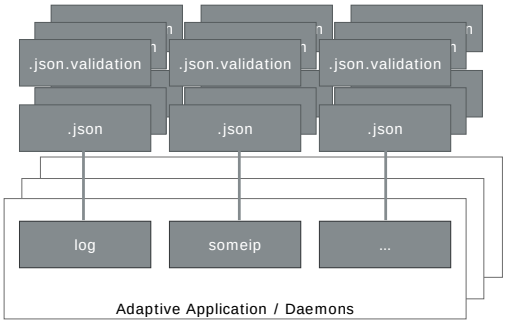


Fig. 1: Every component has its own json files. [2]

Figure 1 illustrates the current architecture, where configuration is managed at the component level. Each component maintains its own configuration files and includes logic to read and interpret them. By introducing a database, this logic could be centralized into a unified "database configuration component." Configuration management can then be organized across different layers.

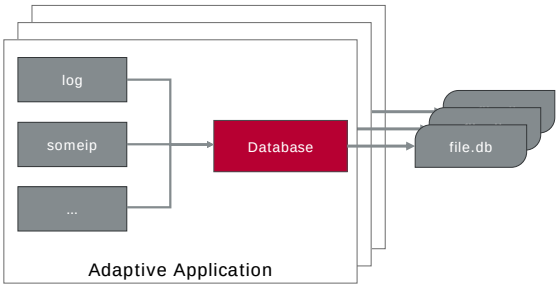


Fig. 2: Configuration is saved on the application level. [2]

Figure 2 presents a concept where configuration data is centralized on a per-application basis. In this approach, any component requiring configuration data interacts with a dedicated configuration component within the application. This configuration component manages all necessary data using a separate, lightweight database file tailored to the application's needs. Each database file is optimized for efficient read operations, ensuring streamlined performance for individual applications.

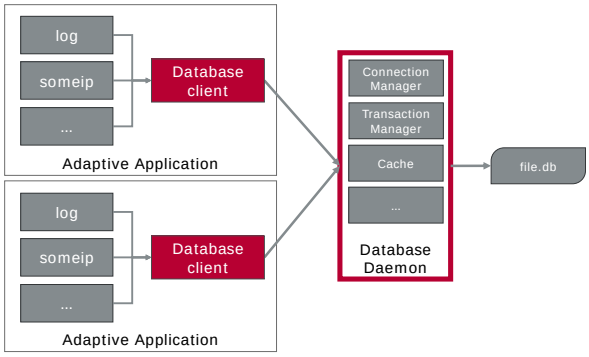


Fig. 3: A daemon manages the configuration in a centralized way. [2]

Figure 3 introduces a concept based on a central system that manages all configuration data. Similar to the application-based approach, a central component within each application handles configuration data. However, in this model, the central component does not directly store the configuration. Instead, it connects to a database-daemon. The daemon manages and serves the required consistency and centralized control.

Evaluation and Outlook

The evaluation of this work is conducted in two key phases. Initially, an analysis of the current system is performed to identify performance bottlenecks and establish a baseline for improvement. This involves methods such as assessing the number and size of JSON files required in a typical deployment and examining the current implementation in detail. Kernel event tracing is conducted to gain deeper insights into system-level bottlenecks, while performance measurements of the current implementation establish a baseline for subsequent experiments.
Following this initial analysis, the gathered data is used to assess and refine the proposed database-based configuration architecture. A proof of concept is developed to validate the feasibility and effectiveness of the new approach, focusing on its ability to address the identified bottlenecks. This prototype is analyzed and compared against the existing JSON-based system to quantify performance gains.
The results of this work provide valuable insights into how database technologies can enhance the performance of configuration data management in MICROSAR Adaptive. The findings will highlight both the advantages and trade-offs of the proposed architectures, paving the way for future optimizations.

References and figures

[1] Vector Informatik GmbH NN. MICROSAR Adaptive. https://www.vector.com/de/de/products/products-a-z/embedded-software/microsar-adaptive/#, 2024.

[2] Own representation.

Evaluierung von Automobilen Kommunikationsprotokollen (UDP/IP, SOME/IP, Eclipse uProtocol) für verteilte Objekterkennungsdienste im Fahrzeug

Pantelis Stefanakos **Dennis Grewe**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Kommunikationsprotokolle als Bindeglied zwischen verteilten Systemen im Fahrzeug

In der Ära der fortschreitenden Vernetzung und Digitalisierung moderner Fahrzeuge hat sich die Kommunikation in verteilten Systemen zu einem komplexen Bereich entwickelt. Kommunikationsprotokolle fungieren als entscheidende Bindeglieder zwischen den verschiedenen Systemen im Fahrzeug, indem sie den Datenaustausch zwischen den Komponenten ermöglichen und koordinieren. Diese Protokolle sind technische Spezifikationen, die das Zusammenspiel der Fahrzeugsysteme steuern. Von der Motorsteuerung bis zum Infotainmentsystem sowie von Sicherheitsfunktionen bis hin zu Fahrerassistenzsystemen muss jede Komponente nahtlos kommunizieren. Das Bordnetz eines modernen Fahrzeugs ist entscheidend für die nahtlose Kommunikation. In einer Limousine umfasst es bis zu 1.500 Einzelleitungen, wiegt rund 50 Kilogramm und integriert etwa 100 Steuergeräte (ECUs) [1].

Das User Datagram Protocol über Internet Protocol (UDP/IP) ermöglicht eine schnelle, verbindungslose Kommunikation zwischen ECUs und ist für zeitkritische Anwendungen mit geringer Latenz geeignet. Das Scalable service-Oriented Middleware over IP (SOME/IP) fördert den effizienten Austausch von Informationen zwischen verschiedenen ECUs und verbessert somit die Interoperabilität komplexer Systeme. Das Eclipse uProtocol ermöglicht eine flexible Datenübertragung zwischen Softwarekomponenten und unterstützt die Integration neuer Technologien; es ist jedoch noch nicht in realen Fahrzeugen implementiert, sondern eine Entwicklung für zukünftige Fahrzeuge im Kontext des Software Defined Vehicle.

Zielsetzung

Im Fokus steht der Anwendungsfall sowie das Hardware-Setup zur Evaluation der Protokolle UDP/IP, SOME/IP und Eclipse uProtocol.

Anwendungsfall: Objekterkennung mit einer Kamera zur Datengenerierung

Um die Kommunikationsprotokolle im Rahmen der Bachelorarbeit zu evaluieren, wird ein spezifischer Anwendungsfall definiert, der praxisnahe Anforderungen an die Datenübertragung stellt. Dabei dient eine Kamera als zentrale Komponente zur Erzeugung von Daten, die innerhalb eines verteilten Systems in einem lokalen Netzwerk übertragen werden sollen. Die von der Kamera aufgenommenen Bilder werden mit einem Objekterkennungsmodell analysiert, das speziell darauf trainiert wurde, vier deutsche Automarkenlogos zu identifizieren: Audi®, BMW®, Mercedes-Benz® und Volkswagen®. Das Modell analysiert die aufgenommenen Bilddaten und ordnet erkannte Logos einer dieser Marken zu. Die Ergebnisse dieser Klassifizierung – also die identifizierten Automarkenlogos – bilden die zu übertragenden Informationen.

Hardware-Setup

Für das Projekt wird ein Hardware-Setup aus zwei Raspberry Pi 5 Modellen, einem Laptop, einem Switch und Ethernet-Kabeln verwendet. Diese Konfiguration bildet ein praxisnahes Testumfeld zur Nachbildung eines verteilten Systems, das zwar nicht direkt in Fahrzeugnetzwerken eingesetzt wird, aber grundlegende Konzepte und Kommunikationsprotokolle simulieren kann, die in modernen Fahrzeugsystemen Anwendung finden. Die beiden Raspberry Pi fungieren als eigenständige Knoten innerhalb des Netzwerks. Einer der Raspberry Pis ist mit einer Kamera verbunden, die über ein Flachbandkabel angeschlossen ist und als

Datenquelle dient. Dieser Raspberry Pi übernimmt die Funktion des Servers. Der zweite Raspberry Pi, der die Daten empfängt, übernimmt die Funktion des Clients. Die Datenübertragung zwischen den beiden Raspberry Pis erfolgt über den Switch.

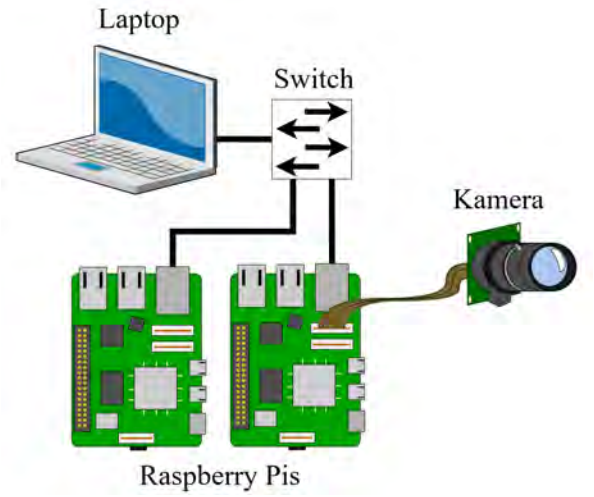


Abb. 1: Netzwerkdiagramm mit Kameramodul [2]

eigenständig erstellt. Die sichtbaren Markenlogos (BMW®, Mercedes-Benz®, Volkswagen®, Audi®) sind geschützte Markenzeichen und Eigentum der jeweiligen Rechteinhaber.

Abb. 2: Visuelle YOLOv8-Ausgabe [2]

Verwendung eines KI-Modells für die Erkennung von Automarken

In dieser Untersuchung wird das YOLOv8-Objekterkennungsmodell (You Only Look Once) von Ultralytics eingesetzt, das speziell für die Erkennung von Automarkenlogos trainiert wurde. Das Hauptziel dieser Untersuchung besteht in der Evaluierung der Kommunikationsprotokolle, wobei die erkannten Logos ausschließlich als Testdaten für die Netzwerkübertragung dienen und auf dem Client-Gerät nicht weiterverarbeitet werden. Das Bild 2 veranschaulicht die visuelle Ausgabe des von Ultralytics entwickelten YOLOv8-Modells. Die Modellausgabe umfasst automatisch generierte Begrenzungsrahmen, die die erkannten Objekte markieren und deren Position sowie Größe angeben. Jedem Rahmen wird ein spezifischer Klassenname zugeordnet, und ein Konfidenzwert quantifiziert die Zuverlässigkeit der jeweiligen Vorhersage [3]. Das ursprüngliche Bildmaterial wurde

Ausblick

Die Bachelorarbeit bietet einen umfassenden Einblick in die Kommunikationsprotokolle UDP/IP, SOME/IP und Eclipse uProtocol. Sie untersucht nicht nur die technischen Aspekte dieser Protokolle, sondern auch die Anforderungen an Entwickler für deren erfolgreiche Implementierung in dem beschriebenen Hardware-Setup. Ein zentraler Bestandteil der Arbeit ist die vergleichende Analyse der Protokolle, die durch Zeitmessungen und eine Bewertung ihrer Eignung für verschiedene Anforderungen in der Automobilindustrie erfolgt. Darüber hinaus beleuchtet die Studie den aktuellen Stand der Technik im Bereich verteilter Systeme in Fahrzeugen. Hierbei werden die wichtigsten Feldbusse, Ethernet-basierte Lösungen und AUTOSAR (AUTomotive Open System ARchitecture) als Schlüsselelemente moderner Fahrzeugarchitekturen vorgestellt. Diese umfassende Betrachtung bietet wertvolle Einblicke in die aktuelle technologische Entwicklung der Automobilelektronik.

Literatur und Abbildungen

[1] Audi AG. Die Assistenzsysteme. https://www.audi-mediacenter.com/de/vorsprung-durch-technik-neu-definiert-der-audi-a8-l-2745/die-assistenzsysteme-2803, 2017.

[2] Eigene Darstellung.

[3] Glenn Jocher et al. Object Detection. https://docs.ultralytics.com/tasks/detect/, 2023.

Sensitivitätsanalyse eines Notbremsassistenten für Nutzfahrzeuge – ein Beitrag zur funktionalen Sicherheit

Leon Struck **Reiner Marchthaler**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Daimler Truck AG, Untertürkheim

Einleitung

In der heutigen Welt gewinnen Advanced Driver Assistance Systems (ADAS) immer mehr an Bedeutung für die Gewährleistung eines möglichst sicheren Straßenverkehrs. Sowohl Kunden als auch Gesetzesgeber verlangen den Einbau von vertrauenswürdigen Fahrerassistenzsystemen. Ein großer Antrieb hierfür ist der Pakt „Vision Zero“ zwischen der Europäischen Union und den Fahrzeugherstellern, dem das Bestreben zugrunde liegt, die Zahl der Unfalltoten im europäischen Gesamtraum schrittweise auf null zu reduzieren. Um dieses Ziel zu erreichen ist es essenziell sicherzustellen, dass die Funktionalitäten der Fahrerassistenzsysteme eine genügende Robustheit gegen Einflüsse und Fehler aufweisen. Die Grundlage hierfür bildet die ISO 26262, die sich mit funktionaler Sicherheit, also der Abwesenheit von unzumutbaren Risiken, der Elektrik und/oder Elektronik (E/E) im Bereich der Straßenfahrzeuge befasst. [3] [5]

Aufgabenstellung

Das Ziel dieser Arbeit ist die Untersuchung eines aktiven Notbremsassistenten auf, durch fehlerhafte Eingangsdaten verursachtes, gefahrbringendes Fehlverhalten. Diese Sensitivitätsanalyse soll hierdurch auch Fehler mit besonders kritischen Folgen bezüglich der Systemreaktion identifizieren. Des Weiteren soll auch speziell das Kriterium zur Erlaubnis eines Notbremsmanövers analysiert werden. Dieses stellt sicher, dass das Ziel der funktionalen Sicherheit, keine ungerechtfertigte Notbremsung auszulösen, eingehalten wird.

Messfahrten

Bereits zu Beginn der Arbeit stehen einige Messfahrten zur Analyse des Systems bereit. Hierbei fährt das „Vehicle under Test“ (VUT), für diesen Test auch Ego-Fahrzeug genannt, auf ein standardisiertes, in Abbildung 1 zu sehendes, „Global Vehicle Target” (GVT) zu. Dieses ist entsprechend der ISO 19206-3 konform, verfügt über einen Zweifrequenzempfänger für globale Navigationssatellitensysteme (GNSS) und repliziert die Eigenschaften eines typischen PKWs gegenüber optischen Sensoren, sowie Radar- und Lidarsensoren.

Abb. 1: Global Vehicle Target (GVT) [2]

Um die zuvor genannten fehlerhaften oder ungenauen Eingangsdaten zu simulieren, werden im Ego-Fahrzeug verschiedene Fehler auf die internen Signale aufgeschalten. Die Tests lassen sich in zehn verschiedene Fälle mit jeweils verschiedenen Geschwindigkeiten, Beschleunigungen und Verhaltensweisen des Ego-Fahrzeugs und des GVTs unterteilen. Jeder dieser Testfälle wird jeweils mehrfach mit einer Reihe an unterschiedlichen Fehleraufschaltungen durchgeführt. Unter diesen Fehlern befinden sich beispielsweise Manipulationen der Signale der Geschwindigkeit, Beschleunigung oder Gierrate des VUT. Da die manipulierten Signale gleichzeitig auch Eingangssignale der umgebungserfassenden Sensortechnik sind, erlaubt dieses Vorgehen ebenfalls eine Betrachtung der Fehlerfortpflanzung. Besonders von Bedeutung sind hierbei die Auswirkungen auf die Reaktion des Systems. [4]

Vorgehensweise

Die Auswertung der Testfahrten läuft hauptsächlich in zwei Softwareumgebungen ab. Die Erste hiervon ist das Tool CANape mit der zusätzlich erweiternden Option Driver Assistance. Hier können fahrzeuginterne Signale aus einzelnen Tests im Nachhinein genau betrachtet

werden. Besonders nützlich ist hierfür das integrierte Szenenfenster. Es ermöglicht eine dynamische grafische Darstellung der Objekte, die von der Sensortechnik des Fahrzeugs aufgenommen werden. Zusammen mit einem nebenliegenden Videofenster, das eine Aufnahme aus dem Testequipment abspielen kann, lässt sich das Gesamtgeschehen in Echtzeit mit der aktuellen Datenlage vergleichen. Die zweite Softwareumgebung ist eine Gruppe von Skripten in MATLAB, die im Rahmen dieser Arbeit entwickelt werden. Diese sollen durch mehrere Funktionen einen breiteren Überblick über die verschiedenen aufgeschalteten Fehler und deren Auswirkungen auf die Sensordaten bieten und die Möglichkeit zu Vergleichen bieten. Durch Eingabe der Bezeichnung der gewünschten Testfälle werden die entsprechenden Dateien automatisch eingelesen und verarbeitet. Die Identifikation der relevanten Objekte für diese Tests geschieht automatisch. Im nächsten Abschnitt werden für die einzelnen Sensorobjekte verschiedene Metriken berechnet, die zur Evaluation der Größe der Abweichung zu den aufgenommenen Daten aus dem Differential GPS des GVT, welche hier als Ground Truth angesehen werden. Daraufhin werden Graphen generiert, die es ermöglichen sollen, bestimmte Signale und Daten aus den verschiedenen Testfällen zu vergleichen. So wird beispielsweise der Fahrweg des GVT relativ zum Ego-Fahrzeug dargestellt, indem die gemessenen Abstände in longitudinale und laterale Richtung übereinander geplottet werden. Als Orientierung werden ebenfalls die korrespondierenden Daten aus dem GVT eingeblendet. Ebenfalls dargestellt werden die, im vorherigen Schritt berechneten, Abstandsfehler zwischen den gemessenen Objekten und der Position des GVT. Diese Art von Graph ist besonders nützlich, um die Abweichungen zum korrekten Wert im Detail zu betrachten. Eine vereinfachte Darstellung des Ablaufs dieses Skripts ist in Abbildung 2 zu sehen.

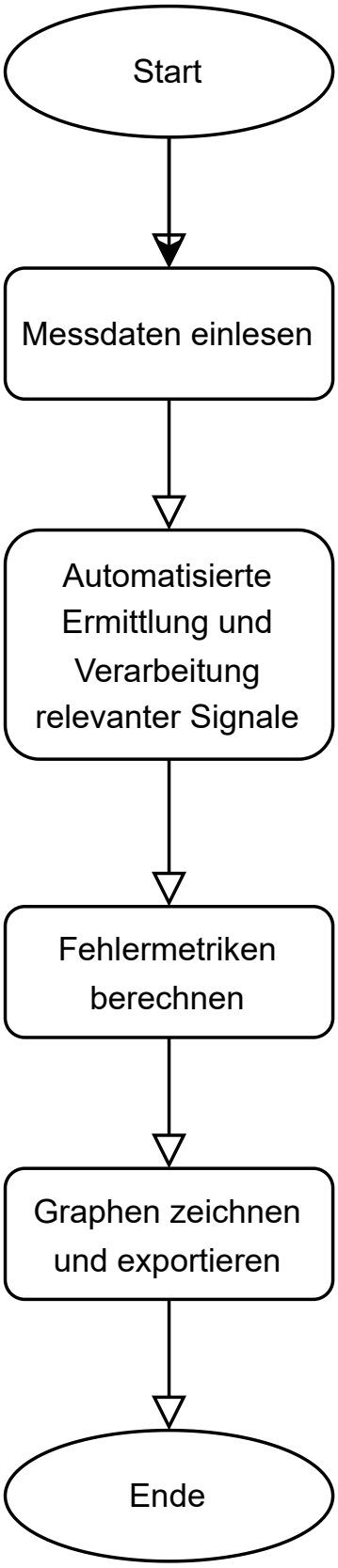


Abb. 2: Programmablaufplan der Fehlerberechnung und Graphenerstellung [1]

Die Evaluierung des Erlaubniskriteriums, das für das Auslösen eines Notbremsmanövers benötigt wird, erfolgt auf mathematischem Wege. Hierbei werden die Eingangsvariablen des Kriteriums über verschiedene Zahlenbereiche variiert. Das Ergebnis des Kriteriums kann zum Schluss über die verschiedenen Variablen und deren Variation geplottet werden, was Rückschlüsse über die jeweiligen Einflüsse auf das Resultat liefert. So sollen Szenarios identifiziert werden, in denen eine unberechtigte Notbremsung freigegeben werden könnte, da dies eine Gefahr für nachfolgende Verkehrsteilnehmer darstellen würde und somit gegen die Prinzipien der funktionalen Sicherheit verstößt.

Ausblick

Die Sensitivitätsanalyse bietet ein aussagekräftiges Maß für die Robustheit des hier behandelten Notbremsassistenten, die sowohl auf mathematischen Grundlagen als auch realitätsbezogenen Beobachtungen beruht. Hierdurch sollten sich auch zukünftig Schlüsse über die Kritikalität eines Fehlers ziehen lassen. Besonders die Betrachtung der reinen Sensordaten könnte eine nützliche Grundlage für die Absicherung zukünftiger Systeme bieten.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] The European New Car Assessment Programme. Collision Avoidance Frontal Collisions Truck-to-Vehicle Test Protocol. https://www.euroncap.com/media/80745/euro-ncap-trucks-ca-frontal-collisions-vehicle-test-protocol-v10.pdf, 05 2024.

[3] Climate Infrastructure and European Environment Executive Agency. EU Road Safety: Towards.

[4] International Organization for Standardization. *ISO 19206:2021 Road vehicles — Test devices for target vehicles, vulnerable road users and other objects, for assessment of active safety functions*. International Organization for Standardization, 2021.

[5] International Organization for Standardization. *ISO 26262:2018 Road vehicles — Functional safety*. International Organization for Standardization, 2018.

Konzept zur Identifikation von Fehlerursachen bei der Inbetriebnahme einer Diagnosetoolkette bei einem Automobilhersteller

Silas Supke **Mirko Sonntag**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Dr. Ing. h.c. F. Porsche AG, Weissach

Einleitung

Sicherheit, Komfort und Energieeffizienz stehen im Fokus moderner Fahrzeuge. Die Lösung dazu liegt in vernetzten Systemen, mit stetig wachsenden Elektronik- und Softwareanteilen. Mit der dadurch steigenden Komplexität sind Standards, neue Technologien und geeignete Softwarelösungen unerlässlich. Eine zentrale Rolle übernimmt dabei die Fahrzeugdiagnose – der Schlüssel zur Bewältigung dieser Herausforderungen [2]. Die Fahrzeugdiagnose beschäftigt sich mit der (Diagnose-)Kommunikation über Diagnoseprotokolle. Dazu gehört das Updaten, Konfigurieren, Inbetriebnehmen und Prüfen der Steuergeräte im Fahrzeug. Die für diese Fahrzeugdiagnose von dem Entwicklungsingenieur oder dem Werkstadttechniker benutzten Softwarewerkzeuge (Diagnosetools) sind die Anwendungen einer komplexen Kette von zusammengesetzten Komponenten, welche die Diagnosekommunikation für den Nutzer zugänglich macht.

Motivation

Aufgrund der Zusammensetzung der Diagnosetoolkette aus verschiedenen herstellerunabhängigen und austauschbaren Schichten ist eine fehleranfällige Konfiguration der standardisierten Schnittstellen notwendig. Durch einen regelmäßigen Austausch von Komponenten in der Diagnosetoolkette und den damit verbundenen Installationen besteht außerdem die Gefahr, dass Pfade in den Konfigurationsdateien nicht korrekt aktualisiert oder Umgebungsvariablen falsch gesetzt werden. Um der zeitintensiven Fehlersuche von diesen oder ähnlichen Fehlern entgegenzuwirken, soll ein Konzept entwickelt werden, welches generisch mögliche Fehler in Installation, Konfiguration und Feinabstimmung findet und damit die Handhabung für den Nutzer erleichtert. Das Ziel des Konzeptes ist es nicht, Fehler in den einzelnen ausgereiften Komponenten der Diagnosetoolkette zu finden.

Grundlagen

Die Grundidee ist standardisierte Datenformate und ein Laufzeitsystem für die Implementierung eines Applikation- und Diagnosesystems zu entwerfen [4]. Die Implementierung dieses Systems (siehe Abb. 1) ist der sogenannte MVCI-Server („Modular Vehicle Communication Interface Server“). Dieser ist über die Diagnose-Server API (D-Server API) von der Anwendung erreichbar und gibt die Kommunikation über die D-PDU API (Diagnostic- protocol data unit application programming interface) an die „MVCI-protocol modul software“. Die „MVCI-protocol modul software“ kommuniziert über das „Vehicle Communication Interface“ (VCI) mit dem Fahrzeug. Standardisiert sind diese Schichten und Schnittstellen in der ISO 22900 Norm.

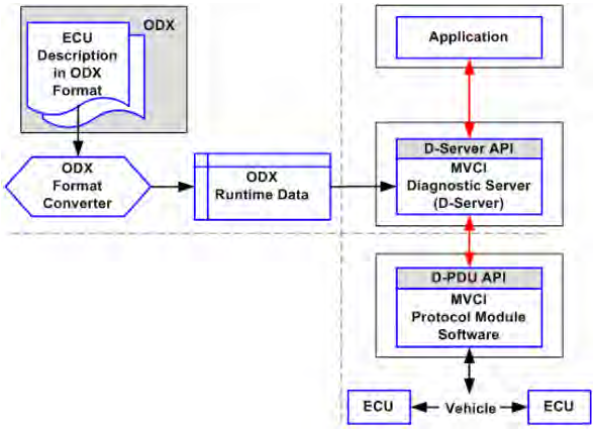


Abb. 1: MVCI-Software Architektur [4]

Der MVCI-Server managt die Datenbank und stellt die notwendigen Informationen für ein einzelnes MVCI-Server Diagnose Server Objekt zur Verfügung. Die Datenbank ist dabei nicht einem bestimmtem Datenbank Objekt zugeordnet, sondern ist im gesamten

Diagnose Server verfügbar [4]. Das Datenmodell „Open Diagnostic Data Exchange“ (ODX) spezifiziert die Haltung bzw. den Austausch aller Diagnoserelevanter Daten. Die hier modellierten Diagnosedaten sind kompatibel mit den Software-Anforderungen des MVCI-Servers. Die ODX-Spezifikation enthält also das Datenmodell zur Beschreibung aller Diagnosedaten des Fahrzeuges [3]. Diese Daten werden in einer indizierten ZIP-Datei (standardisiert als „packed ODX“ oder kurz PDX) im Projekt Ordner abgelegt. Weil hierbei eine geringe Fehleranfälligkeit besteht, wurde die Anbindung der ODX-Daten nicht in das Konzept aufgenommen. Die Anwendung greift auf das „MVCI-Protocol Module“ über die D-PDU API zu. Dabei findet das Ressourcen-Management in der D-PDU API statt. Eine Diagnoseanfrage, die von der Anwendung an die D-PDU API geschickt wird, holt sich das „MVCI-Protocol Module“ aus der D-PDU API und sendet sie durch die Informationen aus dem Diagnoseprotokoll an die Steuergeräte des Fahrzeuges. Durch diese direkte Verbindung zum Fahrzeug ist das „MVCI-Protocol Module“ eine der Schlüssel Komponente, um einen Austausch von Implementierungen der Software und der Diagnose-Protokolle zu gewährleisten.

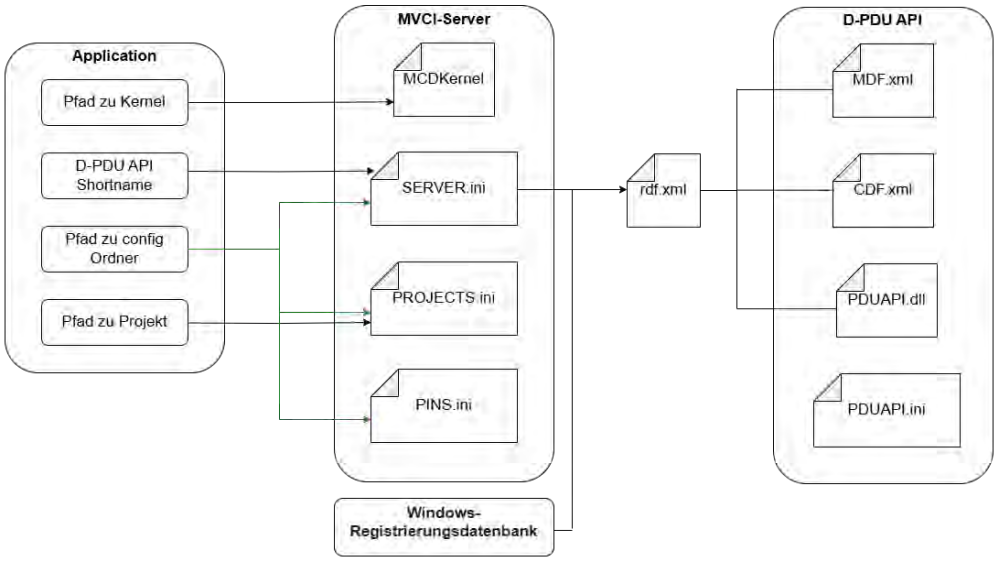


Abb. 2: Datei Abhängigkeiten der Diagnosetoolkette [1]

Ausblick

Bei dem Konzept für die Identifikation von Fehlern, werden die verschiedenen Bereiche, in denen Fehler auftreten können, in drei Schritten abgearbeitet. In Schritt eins wird eine korrekte Installation und eine passende Systemumgebung validiert. Darunter fällt unter anderem eine Überprüfung, ob die richtigen Firewall-Regeln konfiguriert sind und benötigte Ports zu Verfügung stehen. Außerdem findet eine Validierung statt, dass konstant 32- beziehungsweise 64-Bit Versionen installiert wurden und diese Installationen die Umgebungsvariablen, Pfade und Einträge der Windows-Registrierungsdatenbank richtig gesetzt haben. In Schritt zwei werden Datei-Abhängigkeiten (siehe Abb. 2) überprüft und es wird validiert, ob die Konfiguration von dem Nutzer korrekt vorgenommen wurde. Dabei hauptsächlich relevant sind das „root description file“, „module description file“, „cable description file“ und die „Server.ini“. Die „Server.ini“ wird bei Programmstart einmal geladen und es werden hier grundlegende Eigenschaften für den Betrieb des MVCI-Servers konfiguriert. Die RDF-Datei definiert vorhandene Ressourcen in Bezug auf die D-PDU API. Darunter fallen auch die MDF-Datei und die CDF-Datei. Die MDF-Datei beschreibt alle relevanten Rahmenbedingungen für die Diagnosekommunikation über ein „MVCI-Protocol Module“. Die CDF-Datei mappt die Pins des „MVCI-Protocol Module“ auf die Pins der OBD-Buchse (On-Board Diagnose Buchse) des Fahrzeuges. Bei der Server.ini muss unter anderem der Pfad überprüft werden, der auf die RDF-Datei verweist. Sollte dieser Pfad nicht angegeben sein, wird der Default Pfad aus der Windows-Registrierungsdatenbank verwendet. Bei beiden Pfaden muss die Existenz der korrekten RDF-Datei überprüft werden und anschließend mithilfe des „Shortnames“ (Möglichkeit zu Identifikation der einzelnen D-PDU APIs) validiert werden, ob die passende D-PDU API eingetragen ist. In der RDF-Datei sind auch die Pfade zu der MDF-Datei und der CDF-Datei zu überprüfen. Die CDF- und MDF-Datei kann mithilfe von Standards und herstellerspezifischen Pin-Angaben auf Korrektheit überprüft werden. Abschließend kann bei Bedarf eine Detail-Analyse über Abfrage von Status-Meldungen und testen der einzelnen Komponenten der Diagnosetoolkette stattfinden.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Emotive GmbH. User Documentation - VehicleDiagnostic. https://www.emotive.de/wiki/index.php?title=VehicleDiagnostics, 11 2014.

[3] International Organization for Standardization ISO. *ISO 22901-1:2008 Road vehicles – Open diagnostic data exchange – Part 1: Data model specification (ODX)*. Technical Committee ISOTC 22, Road vehicles, Subcommittee SC 3 Electrical and electronic equipment, 2008.

[4] International Organization for Standardization ISO. *ISO 22900-3:2012 Road vehicles - Modular vehicle communication interface (MVCI) - Part 3: Diagnostic server application programming interface (D-Server API)*. Technical Committee ISOTC 22, Road vehicles, Subcommittee SC 3 Electrical and electronic equipment, 2012.

HTMX als leichtgewichtige Alternative für die Webanwendungsentwicklung

Michael Toetsches **Jörg Nitzsche**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma ISTEC Industrielle Software-Technik GmbH, Max-Lang-Str. 56/2 D-70771 Leinfelden-Echterdingen

Einleitung

In der heutigen Zeit stellt die Auswahl des geeigneten Frameworks oder der passenden Bibliothek einen entscheidenden Schritt bei der Entwicklung von Webapplikationen dar. In den letzten Jahren hat sich der Trend zu Single-Page-Applikationen (SPAs) etabliert, wobei Frameworks wie Angular, Vue und die Bibliothek React dominieren. Diese Technologien basieren auf Client-Side Rendering und bringen vorgefertigte Komponenten als Bausteine mit, wobei JavaScript-Code auf der Client-Seite ausgeführt wird, um bei Bedarf dynamisch nachgeladen zu werden. Dies erhöht die Benutzerfreundlichkeit erheblich, da nicht die gesamte Seite neu geladen werden muss. Ein wesentlicher Nachteil dieser Frameworks ist jedoch die steile Lernkurve, die sie aufweisen. Entwickler, die hauptsächlich im Backend tätig sind und über begrenzte Erfahrung in der Webentwicklung verfügen, stehen vor der Herausforderung, sich intensiv in diese umfangreichen Frameworks einzuarbeiten, selbst wenn sie nur gelegentlich Aufgaben im Frontend übernehmen müssen.

Hier kommt Hypertext Markup Extensions (HTMX) ins Spiel. HTMX ist eine browserorientierte JavaScript-Bibliothek, die keine komplexen Abhängigkeiten benötigt und kein Build-System erfordert. Stattdessen ermöglicht sie direkten Zugriff auf moderne Browserfeatures durch die Nutzung von HTML anstelle von JavaScript [2]. Carson Gross, der Entwickler von HTMX, beschreibt die Bibliothek als: Es handelt sich um eine JavaScript-Bibliothek, mit der man Attribute hinzufügen kann, die den href– und action-Attributen von Links und Formularen in Standard-HTML sehr ähnlich sind"[3].

Motivation für eine leichtgewichtige Alternative

Vor dem Hintergrund der beschriebenen Herausforderungen in der Webanwendungsentwicklung verfolgt diese Arbeit das Ziel, alternative Ansätze zu etablieren, die eine effiziente und ressourcenschonende Umsetzung dynamischer Benutzeroberflächen ermöglichen. In diesem Zusammenhang soll HTMX als ein innovativer und leichtgewichtiger Ansatz untersucht werden, der potenziell eine Reduktion der Entwicklungs- und Wartungskomplexität bietet, ohne dabei auf die Umsetzung interaktiver Funktionen zu verzichten.

Die Arbeit strebt an, die Einsatzmöglichkeiten von HTMX in der Webentwicklung zu evaluieren und dessen Potenzial als Alternative zu etablierten Frameworks wie React zu beurteilen. Dabei wird insbesondere geprüft, inwieweit HTMX durch die serverseitige Einbindung von Funktionalitäten über HTML-Attribute eine effektive Lösung für Organisationen mit begrenzten Ressourcen darstellt.

Ein zentraler Fokus liegt auf der Verringerung von Eintrittsbarrieren für Backendentwickler, die gelegentlich Frontend-Aufgaben übernehmen müssen. Es wird untersucht, ob durch den Verzicht auf komplexe JavaScript-Frameworks nicht nur der Schulungsaufwand reduziert, sondern auch die Effizienz der Entwicklungsprozesse gesteigert werden kann.

Darüber hinaus soll analysiert werden, wie HTMX in bestehende serverseitige Technologien integriert werden kann und welche Vorteile sich daraus ergeben. Die Arbeit zielt darauf ab, die Praxistauglichkeit von HTMX zu bewerten und aufzuzeigen, ob grundlegende funktionale Anforderungen im Vergleich zu React beibehalten oder sogar vereinfacht erfüllt werden können.

Durch diese Untersuchung soll die Arbeit einen Beitrag zur Identifikation und Bewertung leichtgewichtiger Alternativen in der Webentwicklung leisten. Die gewonnenen Erkenntnisse könnten insbesondere für

Organisationen mit limitierten personellen oder finanziellen Ressourcen von Nutzen sein und Ansätze für eine effizientere Entwicklung interaktiver Webanwendungen aufzeigen.

Ziel der Arbeit

Im Rahmen dieser Arbeit wird untersucht, inwiefern HTMX als leichtgewichtige Alternative für die Webentwicklung dienen kann. Diese Fragestellung ist insbesondere für die ISTEC von Interesse, da die Technologie es auch Entwicklern mit begrenztem Fachwissen im Bereich der Frontend-Entwicklung ermöglichen könnte, Benutzeroberflächen effizient zu erstellen, ohne tiefgreifende Kenntnisse in Frameworks wie React oder Angular erwerben zu müssen.

Konzept der Arbeit

Zur Validierung dieser These wird ein Feature aus dem internen Projektmanagement-Tool der ISTEC (ISTEC-App) ausgewählt, analysiert und mit HTMX neu implementiert. Die ursprüngliche Implementierung des Features basiert auf React für das Frontend und Quarkus als Backend siehe Abbildung 1.

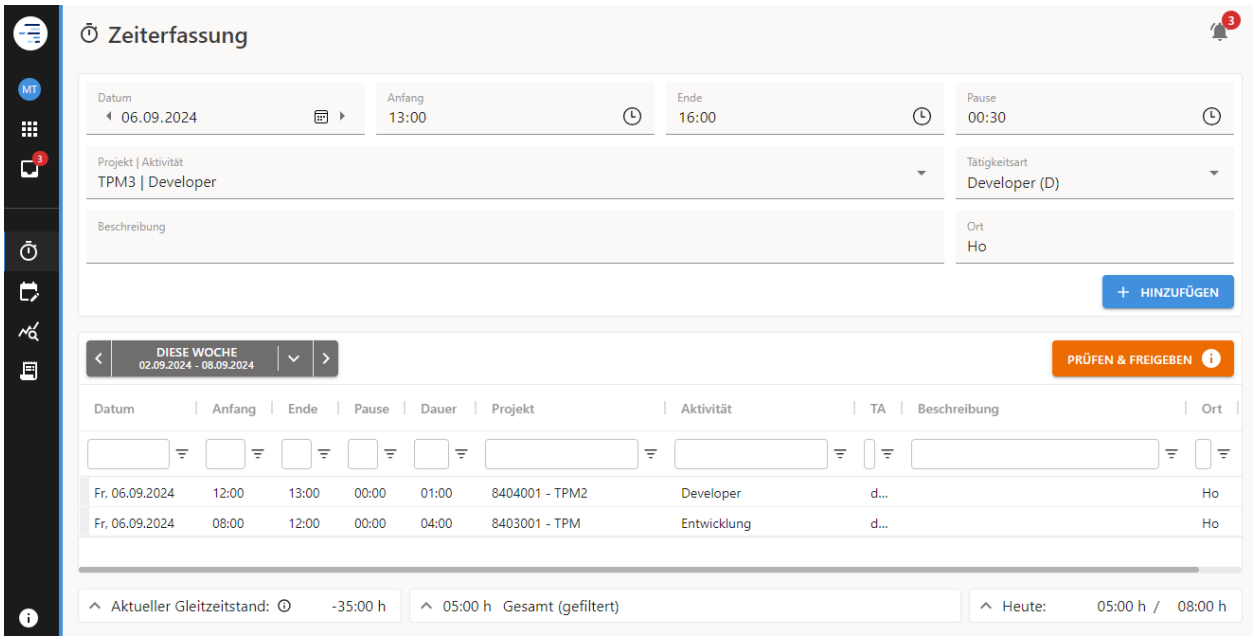


Abb. 1: ISTEC-Applikation: Zeiterfassung Funktion [1]

Im Rahmen der Untersuchung wird die neue Implementierung mit HTMX auf Grundlage von Quarkus mit der bestehenden Lösung verglichen. Der Vergleich erfolgt anhand mehrerer Kriterien: Neben der Code-Komplexität, der Entwicklungszeit und der Lernkurve werden auch die Performanceeigenschaften der beiden Technologien analysiert. Hierbei werden Ladezeiten, Ressourcenverbrauch (CPU, Speicher) und die Reaktionszeit der Benutzeroberfläche gemessen. Zusätzlich wird die fundamentale architektonische Unterscheidung zwischen Client-Side-Rendering (CSR) mit React und Server-Side-Rendering (SSR) mit HTMX betrachtet. Ein weiterer Fokus dieser Arbeit liegt auf der Untersuchung, ob HTMX über seine grundlegenden Einsatzmöglichkeiten hinaus auch für die Realisierung komplexerer Funktionen geeignet ist oder primär auf einfache Benutzeroberflächen beschränkt bleibt. Ziel der Arbeit ist es, auf Grundlage der Ergebnisse eine fundierte Bewertung vorzunehmen, ob HTMX eine adäquate Alternative für den Einsatz in der Webentwicklung darstellen kann und in welchen Szenarien diese Technologie sinnvoll eingesetzt werden kann.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Team HTMX. htmx in a Nutshell. https://htmx.org/docs/, 2024.

[3] Tyson Matthew. HTMX-Schöpfer im Interview: Software ist eine brutale Branche. https://www.computerwoche.de/article/2833120/software-ist-eine-brutale-branche.html#:~:text=HTMX%2DErfinder%20Carson%20Gross%20spricht,Kollegen%20weniger%20sozialer%20Druck%20lastet, 2024.

Entwicklung eines LLM basierten Chatbots für behördliche Bestellvorgänge

Celil Uenal **Mirko Sonntag**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

Die künstliche Intelligenz (KI) gilt heute als eine der wichtigsten Technologien. Nicht nur durch die ständig fortlaufende Innovation und Weiterentwicklung ist diese Technologie so ansprechend, sondern ebenso durch ihre einfache Verfügbarkeit. Man findet KI-Anwendungen mittlerweile in vielen Applikationen und Webanwendungen. Die Entwicklung der KI-Nutzung in der Industrie zeigt in den letzten Jahren einen deutlichen Aufwärtstrend. Laut einer aktuellen Bitkom-Umfrage beschäftigen sich erstmals mehr als die Hälfte der deutschen Unternehmen (57%) mit KI. Der Anteil der Unternehmen, die KI bereits einsetzen, hat sich innerhalb eines Jahres von 15% auf 20% erhöht. Die restlichen 37% planen den Einsatz von KI in naher Zukunft. Dies unterstreicht das Potenzial von KI, die Wettbewerbsfähigkeit deutscher Unternehmen zu stärken und neue Geschäftsmodelle zu ermöglichen. [5]

Large Language Models

Large Language Models (LLMs) sind große Sprachmodelle, die einen wichtigen Aspekt im Themenbereich der generativen KI ausmachen und deren Ziel es ist Texte zu verarbeiten und zu generieren in einer Form, die der natürlichen Sprache entspricht. Die Basis von LLMs bilden neuronale Netze, die durch enorm große Datenmengen trainiert wurden und Textdaten aus Büchern, Artikeln, Webseiten und vielen anderen Quellen umfassen. Ihre Funktionsweise besteht darin, Muster und Zusammenhänge in den Textdaten zu erkennen, zu verstehen, zu analysieren und darauf aufbauende Antworten zu generieren. [1]
LLMs haben ein sehr umfangreiches Potenzial an Einsatzmöglichkeiten, aus Umfangsgründen wird im Folgenden nur eine kleine Auswahl der wichtigsten Punkte betrachtet:

- **Textgenerierung:** Durch das Generieren von Texten können ganze Berichte oder Zusammenfassungen von komplexen Inhalten erstellt werden. Texte können auch leicht in andere Sprachen übersetzt oder kreativ gestaltet werden, z.B. in Form eines Gedichts.
- **Chatbots:** Mithilfe von Chatbots lässt sich jeglicher Support rund um die Uhr gewährleisten. Dadurch lassen sich Anfragen automatisiert und kontextbezogen beantworten.
- **Recherche:** Es ist möglich schnell wichtige Informationen zu beschaffen.
- **Datenanalyse:** Große Datenmengen können mithilfe von LLMs Zusammengeführt und Interpretiert werden, was eine Menge an Zeit ersparen kann.
- **Bildung:** Eine Unterstützung für Lernende ist gegeben, indem komplexe Themen detailliert und verständlich aufbereitet werden.

Problemstellung

Für Bestellungen der Fakultät Informatik und Informationstechnik der Hochschule Esslingen wird eine interne Software zur Bestellverwaltung genutzt. Diese Software dient der Organisation und Übersicht aller Bestellvorgänge. Aktuell ist das System auf die manuelle Bearbeitung von Bestellungen ausgerichtet, was potenziell mit einem hohen Zeitaufwand verbunden ist. Insbesondere die Suche nach passenden Angeboten und die Einhaltung fakultätsinterner Richtlinien, wie die Genehmigungspflicht für Bestellungen über 1.000 €, erfordern zusätzliche Aufmerksamkeit und manuelle Prüfung. Mit dem Ziel, die Bestellprozesse zu automatisieren und zu optimieren, plant die Fakultät die Integration eines Chatbots in das bestehende System. Der Chatbot soll folgende Aufgaben übernehmen, die in Abbildung 1, anhand einer Übersicht bildlich dargestellt werden:

1. Einsicht zu bisherigen Bestellungen: Der Chatbot greift auf die in einer SQL-Datenbank gespeicherten Bestellungen und weitere Daten zu, um dem

Benutzer auf Anfrage relevante Informationen bereitzustellen. Dies soll die Transparenz über vergangene Bestellungen erhöhen und die Nachverfolgbarkeit erleichtern.

2. Unterstützung bei der Angebotssuche: Durch die Integration einer Websuche soll der Chatbot in der Lage sein, passende Angebote für gewünschte Produkte zu finden.

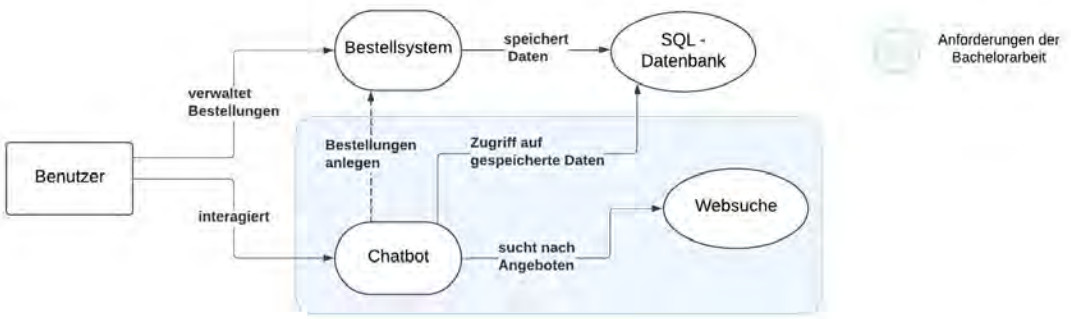


Abb. 1: Übersicht der Anforderungen [2]

Lösungsansätze

SQLAlchemy ist eine beliebte Open-Source Python-Bibliothek, die für die Arbeit mit Datenbanken gut geeignet ist. Dieses Tool kann verwendet werden, um den Zugriff auf die gespeicherten Daten in der SQL-Datenbank für den Chatbot zu ermöglichen. SQLAlchemy hat sich mittlerweile zu einem Standard in der Python-Entwicklung für Datenbankinteraktionen entwickelt. [3]
Web Scraping ist eine Methode zur automatisierten Extraktion von Daten aus Webseiten. Es handelt sich um einen Prozess, bei dem Informationen von Webseiten gesammelt und in ein strukturiertes Format umgewandelt werden, das für weitere Analysen oder Anwendungen genutzt werden kann, dabei wird der Inhalt des HTML-Codes der Webseite extrahiert. Der Einsatz von Web-Scraping-Techniken ermöglicht es, gezielt relevante Angebote zu identifizieren, die den spezifischen Anforderungen der Nutzer entsprechen. [4]

Ausblick

Die Integration eines Chatbots in die Bestellsoftware bietet nicht nur unmittelbare Vorteile wie die Automatisierung von Prozessen und die Verbesserung der Effizienz, sondern eröffnet auch Perspektiven für zukünftige Entwicklungen und Erweiterungen. Langfristig könnte der Chatbot durch die Nutzung von KI und maschinellem Lernen kontinuierlich verbessert werden. Beispielsweise könnte der Chatbot Bestellmuster analysieren, um Vorschläge für zukünftige Einkäufe zu machen oder Nutzer auf günstigere Angebote oder alternative Anbieter hinweisen. Darüber hinaus könnte die Funktionalität des Chatbots erweitert werden, um Genehmigungsprozesse vollständig zu digitalisieren, indem Genehmigungsanfragen automatisch an die zuständigen Personen weitergeleitet werden.

Literatur und Abbildungen

[1] Jan Clusmann, Fiona R. Kolbinger, Hannah Sophie Muti, Zunamys I. Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Loffler, Sophie-Caroline Schwarzkopf, et al. Die kommende Entwicklung großer Sprachmodelle in der Medizin. *Kompass Onkologie*, 11:3–10, 2024.

[2] Eigene Darstellung.

[3] Vinay Kudari. SQLAlchemy — Python Tutorial. https://towardsdatascience.com/sqlalchemy-python-tutorial-79a577141a91, 08 2018.

[4] Heidi Kühnemann. Anwendungen des Web Scraping in der amtlichen Statistik. *AStA Wirtschafts- und Sozialstatistisches Archiv*, 15:5–25, 2021.

[5] Andreas Streim and Janis Hecker. Erstmals beschäftigt sich mehr als die Hälfte der Unternehmen mit KI | Presseinformation | Bitkom e. V. https://www.bitkom.org/Presse/Presseinformation/Erstmals-beschaeftigt-Haelfte-Unternehmen-KI, 2024.

Proaktive Ressourcenskalierung in der Cloud: Ansätze zur Reduktion von Bereitstellungszeiten für virtuelle Rechenkapazitäten im Continuous Software Engineering

Luis Urbitsch **Dieter Morgenroth**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Mercedes-AMG GmbH, Affalterbach

Einleitung

Die jüngste Entwicklung des Automobils hin zum Software-defined Vehicle bietet Herstellern nicht nur unzählige neue Möglichkeiten in der Entwicklung von Sicherheits-, Fahrerassistenz- und Infotainmentsystemen, sondern stellt sie auch vor neue Herausforderungen. Steigende Komplexität, kürzere Entwicklungszyklen und sich stetig verändernde Kundenbedürfnisse erfordern eine Transformation etablierter Methoden, Vorgehensmodelle und Praktiken. Da die Anforderungen an das SDV, wie beispielsweise kontinuierliche Updates von Softwarekomponenten zur Bereitstellung neuer Funktionalitäten, sich verändernde Sicherheitsanforderungen und schnelle Reaktionsfähigkeit auf Marktbedürfnisse, immer stärker denen von Produkten aus der Software- und Unterhaltungselektronikbranche ähneln, adaptieren Automobilhersteller etablierte Ansätze aus diesen Branchen und implementieren sie in ihren eigenen Softwareentwicklungsprozess. Unter anderem können die in der klassischen Softwareentwicklung bereits etablierten Methoden Continuous Integration (CI), Continuous Testing (CT) sowie Continuous Deployment (CD) dazu genutzt werden, um den neu entstandenen Herausforderungen gerecht zu werden [5].

Hintergrund

Das mit diesen Methoden einhergehende kontinuierliche Bauen, Testen und Bereitstellen von Software bringt jedoch auch Anforderungen an das Entwicklungsumfeld und die zugrunde liegende Infrastruktur mit sich. Die in Form von CI-Jobs auftretende Arbeitslast fordert Rechenressourcen und Werkzeuge, die zur Ausführung der verschiedenen Aufgaben benötigt werden. Aufgrund der Heterogenität der anfallenden Arbeitslast, die von einfachen statischen Programmanalysen bis hin zu rechen- und speicherintensiven Simulationen reicht, gestaltet es sich insbesondere in hochkomplexen Entwicklungsumfeldern schwierig, die benötigten Ressourcen aus lokalen Kapazitäten flexibel und skalierbar bereitzustellen. Um den aus dem Continuous Software Engineering erwachsenden Anforderungen an die Infrastruktur gerecht zu werden, können alternativ auch virtuelle Server bzw. virtuelle Maschinen von Cloud-Dienstleistern genutzt werden, um die angefragten Rechenkapazitäten bereitzustellen. In diesem Kontext erweist sich die Möglichkeit, Rechenkapazitäten in Form von virtuellen Maschinen dynamisch für sich verändernde Arbeitslasten bereitzustellen (Autoscaling), als von besonderer Bedeutung.

Herausforderung

Die zentrale Herausforderung besteht darin, die in Abbildung 1 aufgezeigten Verzögerungen zwischen der Erstellung der für einen CI-Job erforderlichen virtuellen Maschine und deren Einsatzbereitschaft zu minimieren.

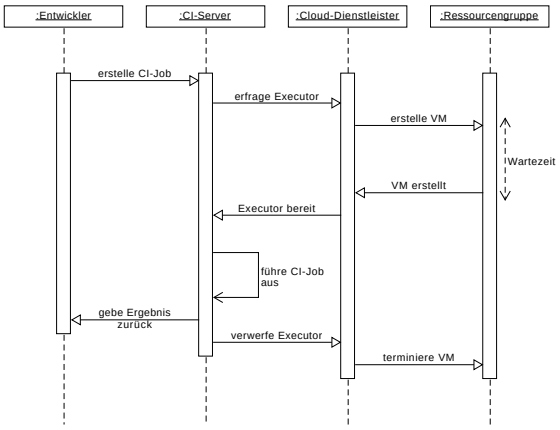


Abb. 1: Ablauf zur Bereitstellung virtueller Rechenkapazitäten für einen CI-Job [1]

Die Bereitstellung virtueller Maschinen in der Cloud kann mehrere Minuten in Anspruch nehmen [2], wodurch sich die Wartezeit bis zur tatsächlichen Ausführung des CI-Jobs erheblich verlängert. Dies beeinträchtigt nicht nur die Effizienz der Entwicklungsprozesse, sondern erschwert auch die zeitnahe Rückmeldung zu Änderungen in der Software an die Entwickler. Diese Problematik ist insbesondere bei kleineren Aufgaben wie statischen Codeanalysen oder Modultests von Bedeutung, da die Wartezeit auf die Bereitstellung der benötigten Ressourcen im Vergleich zur eigentlichen Ausführungsdauer dieser Aufgaben unverhältnismäßig hoch ist.

Zielsetzung und Lösungsansätze

Ziel dieser Arbeit ist es, Ansätze und Strategien zu entwickeln, die eine rechtzeitige Bereitstellung virtueller Rechenkapazitäten für Continuous-Integration-Jobs ermöglichen, um die Wartezeiten zwischen der Anforderung und der tatsächlichen Ausführung zu minimieren. Aufgrund der Latenzzeiten bei der dynamischen Bereitstellung virtueller Maschinen ist es notwendig, Rechenkapazitäten bereits vor ihrem tatsächlichen Bedarf verfügbar zu machen. Ein Ansatz besteht in der Optimierung der statischen Konfiguration der Verwaltungseinheit, die für die Skalierung der Infrastruktur und die Zuweisung von Aufgaben zuständig ist. Dazu gehört beispielsweise, virtuelle Maschinen länger in einem einsatzbereiten Zustand zu halten, sie außerhalb der eigentlichen Nutzung in den Wartebetrieb zu versetzen oder zu definierten Zeitpunkten zusätzliche Kapazitäten vorzuhalten. Die Anwendung dieser Strategien erfordert eine präzise Abwägung, um eine Balance zwischen Verfügbarkeit und Ressourcenkosten zu gewährleisten. Darüber hinaus wird die Nutzung dynamischer Vorhersagemodelle untersucht, die auf historischen Daten basieren, um zukünftige Arbeitslasten vorherzusehen. Diese Modelle sollen es ermöglichen, Rechenressourcen proaktiv bereitzustellen, bevor ein Bedarf entsteht. Arbeiten von Hu et al. [2], Morais et al. [4] sowie von Podolskiy et al. [6] haben bereits gezeigt, dass die Verwendung insbesondere von autoregressiven Vorhersagemodellen (z. B. AutoRegressive Moving Average) gute Ergebnisse bei der proaktiven Bereitstellung von virtuellen Rechenkapazitäten für generische Arbeitslasten liefert.

Vorgehen

Zur Umsetzung und Bewertung der beschriebenen Lösungsansätze wird ein mehrstufiger Ansatz verfolgt. Zunächst ist es notwendig, eine umfangreiche und repräsentative Datenbasis aufzubauen, welche die typischen Anforderungen an Cloud-Infrastrukturen in kontinuierlichen Entwicklungsprozessen abbildet. Hierfür werden Metriken zur Ressourcennutzung und Arbeitslast in Form von CI-Jobs gesammelt und analysiert. Die benötigten Daten für die Umsetzung stammen aus den realen Betriebsprozessen eines Unternehmens, das GitLab CI als Plattform für kontinuierliche Integrations- und Deployment-Prozesse nutzt. Die nachfolgenden Schritte werden auf Grundlage der in diesem spezifischen Anwendungsszenario erhobenen Daten durchgeführt.
Zusätzlich wird eine Simulationsumgebung entwickelt, welche die Komponenten und deren Interaktion im, in Abbildung 2 dargestellten, System nachbildet. Diese Umgebung dient dazu, Experimente durchzuführen, um optimal abgestimmte Konfigurationen zu identifizieren und Vorhersagemodelle auf ihre Eignung für eine präzise und effiziente Ressourcenallokation zu testen. Die Simulationsumgebung wird so entwickelt, dass die Integration und Prüfung verschiedener Vorhersagemodelle möglich ist.

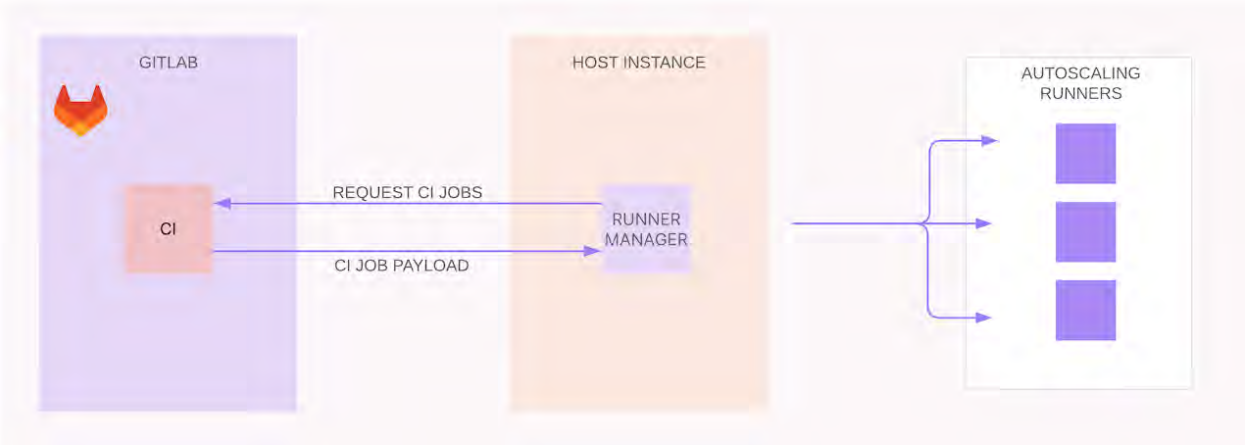


Abb. 2: Komponenten des CI-Systems [3]

Anschließend werden diese Modelle dahingehend analysiert, wie genau sie Arbeitslasten prognostizieren und inwieweit sie zur Optimierung der Ressourcennutzung beitragen können. Der Fokus der Analyse liegt auf der Fähigkeit der Modelle, ein ausgewogenes Verhältnis zwischen Effizienz und Wirtschaftlichkeit zu erreichen. Abschließend sollen die im Rahmen der Simulationsumgebung entwickelten Lösungsansätze in die reale Betriebsumgebung integriert und die Ergebnisse der Simulation mit den tatsächlichen Ergebnissen aus der Praxis verglichen werden, um zu überprüfen, inwieweit sich die gewonnenen Erkenntnisse auf die reale Anwendung übertragen lassen.

Ausblick

Die entwickelten Ansätze zur Reduktion von Wartezeiten fokussieren sich vor allem auf kurzlaufende Jobs, bei denen Bereitstellungsverzögerungen besonders ins Gewicht fallen. Für langlaufende, ressourcenintensive Jobs gewinnen jedoch Faktoren wie die Anzahl der vCPU-Kerne, die Menge an Arbeitsspeicher (RAM), die I/O-Leistung sowie spezifische Optimierungen wie GPU-Unterstützung oder Netzwerkbandbreite an Bedeutung. Zukünftige Arbeiten könnten die Modelle um diese Parameter erweitern, um eine optimale Abstimmung der Maschinenkonfigurationen auf die jeweiligen Anforderungen zu ermöglichen. Darüber hinaus könnten hybride Ansätze, die lokale und Cloud-Ressourcen kombinieren, untersucht werden, um noch höhere Flexibilität und Verfügbarkeit sicherzustellen.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Y. Hu, B. Deng, and F. Peng. Autoscaling prediction models for cloud resource provisioning. In *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*, pages 1364–1369. IEEE, 2016.

[3] GitLab Inc. GitLab Runner Autoscaler. https://docs.gitlab.com/runner/runner_autoscale/#gitlab-runner-autoscaler, 2024.

[4] F.J.A. Morais, F.V. Brasileiro, R.V. Lopes, R. Araujo Santos, W. Satterfield, and L. Rosa. Autoflex: Service Agnostic Auto-scaling Framework for IaaS Deployment Models. In *2013 13th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing*, pages 42–49. IEEE/ACM, 2013.

[5] P. Obergfell, S. Kugele, C. Segler, A. Knoll, and E. Sax. Continuous Software Engineering of Innovative Automotive Functions: An Industrial Perspective. In *2019 IEEE International Conference on Software Architecture Companion (ICSA-C)*, pages 127–128. IEEE, 2019.

[6] V. Podolsykiy, A. Jindal, M. Gerndt, and Y. Oleynik. Forecasting Models for Self-Adaptive Cloud Applications: A Comparative Study. In *2018 IEEE 12th International Conference on Self-Adaptive and Self-Organizing Systems (SASO)*, pages 40–49. IEEE, 2018.

Künstliche Intelligenz in Business Intelligence: Entscheidungsoptimierung im E-Commerce durch Predictive Analytics

Cem Varan **Astrid Beck**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

Es ist für jedes Unternehmen nahezu unverzichtbar geworden, große Datenmengen zu sammeln und zu analysieren. Datenbasierte und innovative Technologien haben insbesondere im E-Commerce, einem schnell wachsenden Wirtschaftszweig, eine zunehmende Bedeutung für den Unternehmenserfolg. Die Entwicklungen dieser datenbasierten Technologien werden maßgeblich von Business Intelligence (BI) und Künstlicher Intelligenz (KI) getragen. Die Integration von KI ermöglicht die Optimierung prädiktiver Analysen, die den strategischen Handlungsspielraum der Unternehmen erheblich erweitern, während traditionelle-BI-Methoden in der Vergangenheit vor allem deskriptive Analysen verwendeten. Predictive Analytics, ein Teilbereich der KI, ist zu einer entscheidenden Technologie für die Optimierung von Entscheidungen geworden, da es erlaubt, auf der Grundlage historischer Daten, künftige Ereignisse und potenzielle Trends vorherzusagen. Diese Methode ist besonders im Bereich des E-Commerce von Bedeutung, in dem sich Kundenanforderungen schnell ändern und der Wettbewerb durch personalisierte Angebote und effiziente Prozesse dominiert wird. [1]

Theoretische Grundlagen

Künstliche Intelligenz (KI): Die KI hat sich von einem damals kleinen Forschungsgebiet zu einem umfangreichen Sektor entwickelt, der durch die schnellen Entwicklungen in der Technologie exponentiell weiterwächst. Vor allem nach der Covid-19-Pandemie boomte der KI-Sektor, und die Pandemie wirkte wie ein Katalysator, der dafür sorgte, dass KI weitgehend genutzt wird und das Interesse an KI-Technologien deutlich anstieg. Die KI zielt darauf ab, nicht nur menschliche Intelligenz zu verstehen, sondern sie wird auch dafür genutzt, um intelligente Maschinen zu entwickeln, die in der Lage sind, eigenständig zu entscheiden und zu handeln, insbesondere in neuen und unvorhergesehenen Situationen. KI ist eine interdisziplinäre Wissenschaft, die sich auf Algorithmen und Systeme konzentriert, die aus Daten lernen und weiterfolgende Handlungen auf dieser Grundlage automatisieren. [3] Zentrale Ansätze der KI im Kontext von BI sind: *Machine Learning (ML)* - Der Kern von KI-basierten Anwendungen ist das ML, das Algorithmen verwendet, um aus historischen Daten zu lernen und neue Muster in Datenmengen zu erkennen. *Deep Learning (DL)* - Eine Weiterentwicklung des ML, bei der neuronale Netze eingesetzt werden. Diese Technologie hat Anwendungen in der Bild- und Videoerkennung und in der Sprachverarbeitung. [5]

Business Intelligence (BI): Business Intelligence ist ein umfassender Ansatz zur Datensammlung, Analyse und Präsentation von Informationen, um dann in der Entscheidungsfindung zu unterstützen, vor allem im Top-Management. Viele Unternehmen sind im Bereich der BI oft noch unzeitgemäß und kommen mit der Menge an Daten nicht mehr hinterher, wie erwartet wurde. [5] Die Menge an Daten verläuft, aufgrund der Weiterentwicklung des Internets und digitalen Produkten und Dienstleistungen, exponentiell, wie man in Abb. 1 sehen kann.

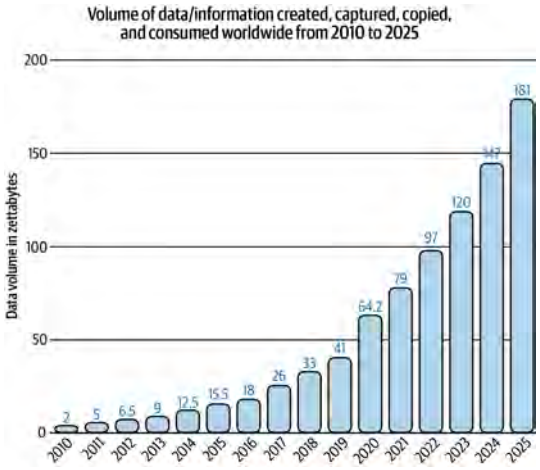


Abb. 1: Datenvolumen weltweit 2010-2025 [5]

Predictive Analytics: Predictive Analytics, als zentraler Bestandteil moderner BI-Systeme, nutzt statistische Modelle und ML-Algorithmen, um zukünftige Trends und Ereignisse vorherzusagen (siehe Abb. 2). Traditionell werden BI-Methoden auf vergangenheitsbezogene Analysen beschränkt, wohingegen KI hier unterstützen kann, um BI-Methoden auf prädiktive Analysen zu erweitern. Somit wird es Unternehmen ermöglicht, datenbasierte Entscheidungsfindung auf ein neues Niveau zu heben, indem sie nicht nur vergangene Trends analysieren, sondern diese auch in konkrete Handlungsempfehlungen für die Zukunft umsetzen.

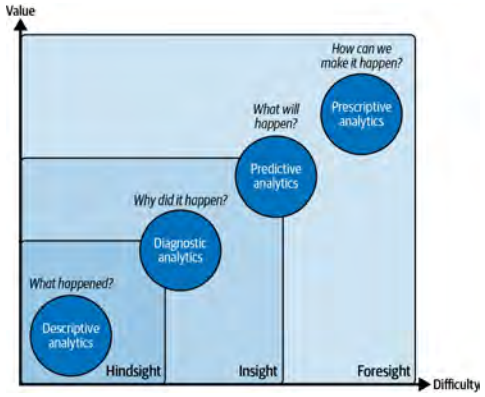


Abb. 2: Methoden der Datenanalyse [5]

Anwendungen im E-Commerce

Die Personalisierung ist ein zentraler Erfolgsfaktor im E-Commerce. Durch Predictive Analytics können individuelle Kundenvorlieben analysiert und gezielt angesprochen werden, was zu einer erhöhten Kundenzufriedenheit und Steigerung der Conversion-Rate beisteuert. [5] *Beispiel:* Amazon nutzt personalisierte Empfehlungen, die durch Collaborative Filtering und ML unterstützt werden, um die Kauferfahrung zu verbessern und den Umsatz zu maximieren. [2] Dynamische Preisgestaltung ermöglicht es Unternehmen, ihre Preise flexibel an Markt- und Nachfragebedingungen anzupassen. Predictive Analytics analysiert historische Verkaufsdaten, aktuelle Markttrends und Wettbewerbspreise, um optimale Preisstrategien zu entwickeln. *Beispiel:* Amazon analysiert kontinuierlich Daten wie Nachfrage, Lagerbestände und Konkurrenzpreise, um mithilfe von Machine-Learning-Algorithmen die Preise in Echtzeit anzupassen. So werden optimale Preise berechnet, die den Umsatz maximieren und gleichzeitig eine effiziente Lagerverwaltung ermöglichen. Diese Strategie ermöglicht es Amazon, wettbewerbsfähig zu bleiben und die Kundenzufriedenheit durch personalisierte Angebote zu steigern. [2]

Herausforderungen

- **Datenqualität und Integration**: Der Erfolg von Predictive Analytics hängt von der Qualität und Integration der zugrunde liegenden Daten ab, d.h. eine saubere und konsistente Datenbasis ist entscheidend für die Genauigkeit von Prognosemodellen. [5]
- **Technologische Anforderungen**: Die Implementierung von Predictive Analytics erfordert leistungsstarke IT-Infrastrukturen und spezialisiertes Fachwissen. Besonders kleine Unternehmen könnten mit den Anforderungen an Hardware und Software überfordert sein. [4]
- **Ethische und rechtliche Aspekte**: Der Umgang mit sensiblen Kundendaten erfordert strenge Einhaltung gesetzlicher Vorgaben. Die Notwendigkeit transparenter und verantwortungsvoller Datenverarbeitung ist wichtig, um das Vertrauen der Kunden zu wahren. [5]

Ausblick

Der Einsatz von Predictive Analytics wird in den kommenden Jahren durch technologische Fortschritte wie Echtzeitdatenverarbeitung, Reinforcement Learning und verbesserte KI-Modelle weiter an Bedeutung gewinnen. Diese Technologien ermöglichen es, Vorhersagen nicht nur auf historischen Daten, sondern auch auf Echtzeitinformationen und externen Faktoren wie globalen Markttrends oder saisonalen Veränderungen zu basieren. Dies wird die Präzision und Geschwindigkeit, mit der Entscheidungen getroffen werden, erheblich steigern. Insbesondere in Bereichen wie der dynamischen Preisgestaltung, der Lagerverwaltung oder der Personalisierung von Angeboten könnten neue Ansätze eine noch höhere Flexibilität und Anpassungsfähigkeit bieten. Unternehmen könnten dadurch besser

auf plötzliche Nachfrageschwankungen reagieren oder kundenspezifische Präferenzen in größerem Umfang berücksichtigen. Perspektivisch könnten auch ethische und nachhaltige Dimensionen stärker integriert werden, beispielsweise durch Modelle, die nicht nur auf Gewinnmaximierung abzielen, sondern auch ökologische und soziale Auswirkungen einbeziehen. Insgesamt eröffnet der fortschreitende Einsatz von Predictive Analytics die Möglichkeit, Geschäftsprozesse datengetriebener, präziser und gleichzeitig umfassender zu gestalten.

Literatur und Abbildungen

[1] Mark Harwardt and Maximilian Köhler. *Künstliche Intelligenz entlang der Customer Journey - Einsatzpotenziale von KI im E-Commerce*. Springer Gabler Wiesbaden, 2023.

[2] Paul Roetzer and Mike Kaput. *Marketing Artificial Intelligence: AI, Marketing, and the Future of Business*. Matt Holt, 2022.

[3] Stuart Russel and Peter Norvig. *Artificial Intelligence - A Modern Approach (Global Edition)*. Pearson Education Limited, 4 edition, 2021.

[4] Ramesh Sharda, Dursun Delen, and Efraim Turban. *Analytics, Data Science, & Artificial Intelligence: Systems for Decision Support, Global Edition*. Pearson, 11 edition, 2020.

[5] Tobias Zwingmann. *AI-Powered Business Intelligence - Improving Forecasts and Decision Making with Machine Learning*. O'Reilly Media, 2022.

Lean ERP-Einführung bei der Mercedes-Benz AG: Entwicklung eines effizienten Rollout-Konzepts in der Logistik

Laura Viscardi **Thomas Rodach**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Mercedes-Benz AG, Esslingen am Neckar

Problemstellung

Die Implementierung des IT-Systems AmSupply spielt eine zentrale Rolle in der Logistik der Mercedes-Benz AG. Dabei steht die Logistik vor der Herausforderung, das IT-System sowohl an neuen also auch an bestehenden Standorten auszurollen. Technische und organisatorische Hürden, insbesondere bei der Integration in bestehende Prozesse, führen oft zu Verzögerungen oder Mehraufwand, die durch ein schlankes Konzept umgangen werden können. Diese Arbeit untersucht, wie IT-System-Rollouts und Systemwechsel bei der Mercedes-Benz AG (insbesondere in der Logistik) effizient gestaltet werden können, um Herausforderungen zu minimieren und Synergien zu nutzen.

Zielsetzung

Ziel der Bachelorarbeit ist es, einen praxisnahen Implementierungsleitfaden zu entwickeln, der auf den Erfahrungen aus dem laufenden Rollout am Standort Kecskemét (Ungarn) basiert. Es soll untersucht werden, wie der Rollout des IT-Systems AmSupply in der Logistik der Mercedes-Benz AG möglichst effizient und reibungslos gestaltet werden kann – sowohl an neuen als auch an bestehenden Standorten. Dabei sollen sowohl technische als auch organisatorische Aspekte beleuchtet werden, um daraus praxisrelevante Handlungsempfehlungen abzuleiten. Der Fokus liegt darauf, zukünftige Rollouts durch standardisierte Ansätze und praxiserprobte Empfehlungen zu optimieren. Abbildung 1 zeigt die wichtigsten Meilensteine des Rollouts, beginnend mit dem Kickoff bis hin zum Projektabschluss.

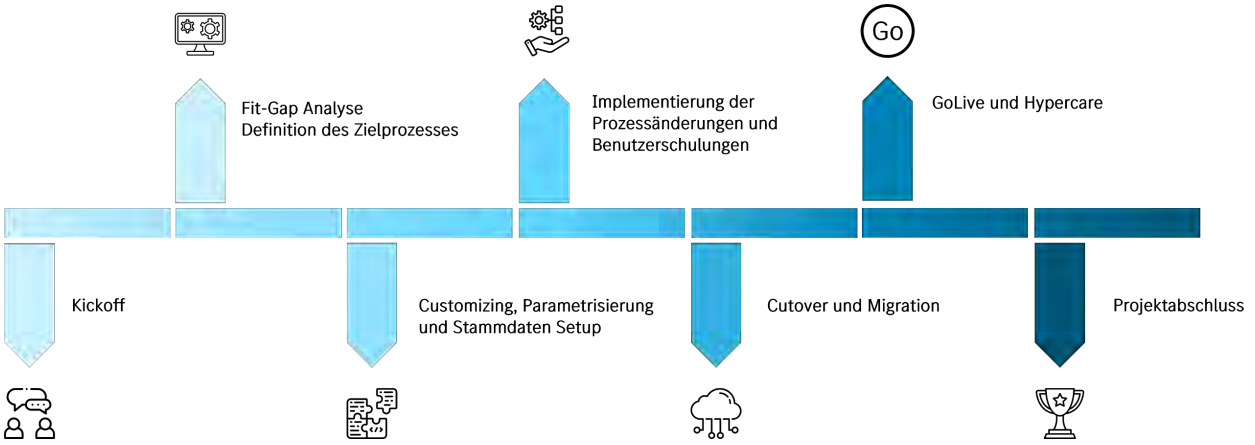


Abb. 1: Meilensteine des AmSupply-Rollouts [2]

AmSupply: Ein SAP-basiertes Template

Das SAP-Logistik-Template Automotive Supply (Abk. AmSupply oder AmS) ist ein spezialisiertes SAP ERP-System, das sämtliche Geschäftsprozesse der Intralogistik innerhalb eines Produktionswerkes abdeckt. Dabei bildet es die Grundlage für die Steuerung und Optimierung von Prozessen und Materialströmen,

die für eine reibungslose Produktionsversorgung erforderlich sind. Zu den zentralen Funktionen gehören die Verwaltung von Materialstammdaten, Verbrauchs- und Bedarfsplanung sowie der Versand von Teilen und Komponenten. Ergänzt wird dies durch Schnittstellen zu externen Partnern wie Zulieferern und Logistikdienstleistern, wodurch ein reibungsloser Datenaustausch ermöglicht wird. Ein SAP Template ist eine standardisierte Vorlage, die global definierte Geschäftsprozesse eines Unternehmens in einem SAP-System abbildet. Diese Templates werden als Grundlage für die Einführung von SAP-Systemen an verschiedenen Standorten oder in unterschiedlichen Geschäftseinheiten eines Unternehmens verwendet. Das Ziel ist es dabei, Prozesse zu harmonisieren, Kosten zu reduzieren und eine einheitliche IT-Systemlandschaft zu schaffen. Das Besondere an SAP Templates wie AmSupply ist die Möglichkeit zur sogenannten „Template-Lokalisierung“ [3]. Während die Hauptfunktionen standardisiert und global gültig sind, können spezifische Anforderungen oder lokale Begebenheiten durch zusätzliche Anpassungen berücksichtigt werden. Diese Flexibilität gewährleistet einerseits die Einhaltung globaler Standards und ermöglicht andererseits die Adaption an individuelle Standortanforderungen. Zusätzlich basiert AmSupply auf einem zentralen Mastertemplate, das die übergeordneten, für alle Werke gültigen Prozesse definiert. Lokale Templates entstehen durch gezielte Anpassungen des Mastertemplates, die spezifische Gegebenheiten eines Standorts berücksichtigen, ohne die globale Einheitlichkeit zu gefährden. Die Systemarchitektur von AmSupply ist in mehrere Ebenen unterteilt, um die Entwicklung, Qualitätssicherung und den operativen Betrieb effizient zu gestalten. Die Entwicklungsumgebung dient als zentrales System, in dem neue Funktionen entwickelt werden. Sie bildet die gemeinsame Basis und wird von allen Werken genutzt. Anschließend werden die neuen Funktionen im Qualitätssystem umfassend getestet, um deren Stabilität und Praxistauglichkeit sicherzustellen. Ergänzend dazu werden spezifische Tests in Testsystemen durchgeführt, um Standortanpassungen und Lokalisierungen zu überprüfen. Nach erfolgreicher Validierung wird die finale Version in die Produktivsysteme integriert, die die operative Grundlage für den laufenden Betrieb bilden und standortspezifische Anpassungen beinhalten. Dieses strukturierte Modell gewährleistet einen reibungslosen und sicheren Ablauf des Rollout- und Betriebsprozesses. Abbildung 2 zeigt eine schematische Darstellung der Systemarchitektur von AmSupply.

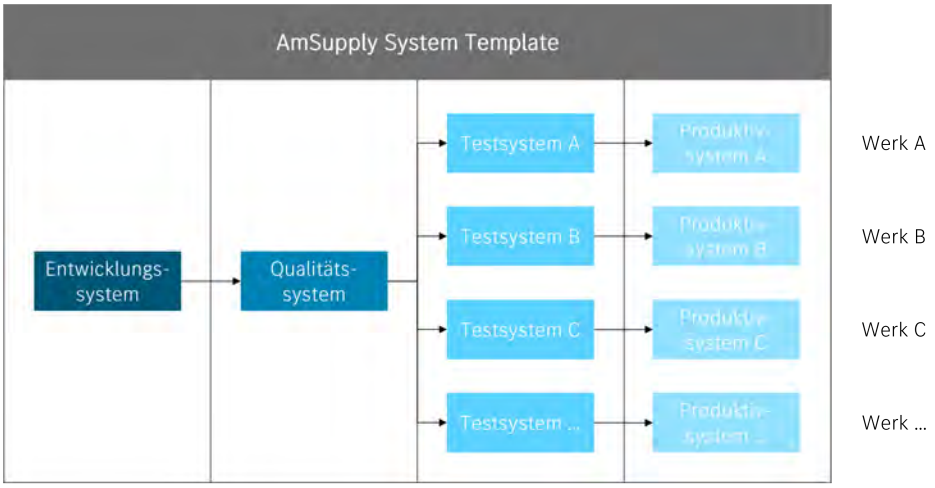


Abb. 2: Systemarchitektur von AmSupply [1]

Ausblick

Die Ergebnisse dieser Arbeit sollen dazu beitragen, die Implementierungsstrategien für AmSupply grundlegend zu verbessern. Der entwickelte Leitfaden dient nicht nur als Grundlage für die Standardisierung künftiger Rollouts, sondern zeigt auch, wie lokale Anpassungen effizient integriert werden können. Langfristig soll der Leitfaden dazu beitragen, die Digitalisierung und Prozessharmonisierung innerhalb der Mercedes-Benz AG weiter voranzutreiben, um Wettbewerbsvorteile durch höhere Effizienz und Flexibilität zu sichern.

Literatur und Abbildungen

[1] Mercedes-Benz AG. Präsentation zur AmSupply Vorstellung (Internes Dokument, nicht öffentlich zugänglich). -, 2023.

[2] Mercedes-Benz AG. Präsentation zum AmSupply Kecskemét Rollout (Internes Dokument, nicht öffentlich zugänglich). -, 2024.

[3] Lukas Bächle. Erfolgreich Standardisieren bei SAP Rollouts. https://www.valantic.com/de/blog/erfolgreich-standardisieren-bei-sap-rollouts/, 2022.

Entwicklung einer Softwarelösung zur Berechnung eines Product Carbon Footprints von Blechkomponenten basierend auf Betriebsdaten komplexer mechatronischer Systeme

Marius Wieler **Mirko Sonntag**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma TRUMPF SE + Co. KG, Ditzingen

Einleitung

Klimaschutz ist unumstritten eines der größten globalen Probleme im 21. Jahrhundert. Ein entscheidender Schritt im Klimaschutz war das Pariser Klimaabkommen von 2015. Dabei verpflichteten sich die Unterzeichnerstaaten einschließlich Deutschland, Maßnahmen einzuleiten, um die Erderwärmung auf maximal 2 Grad Celsius zu begrenzen. In diesem Zusammenhang gewinnt in Deutschland Nachhaltigkeit im Unternehmenskontext zunehmend an Bedeutung, da Unternehmen verstärkt Verantwortung für ihre Umweltauswirkungen übernehmen müssen. Verschiedene Nachhaltigkeitstreiber wie gesetzliche Vorgaben, technologische Innovationen und das gesteigerte Bewusstsein der Verbraucher tragen zu diesem Wandel bei. Trumpf entwickelt Laser für die Blechbearbeitung und ist eng mit der Stahlindustrie verbunden, da Blech unter anderem aus Stahl hergestellt wird. Angesichts der Tatsache, dass die Stahlindustrie in Deutschland etwa 30 Prozent [3] der Treibhausgasemissionen der gesamten Industrie verursacht, richtet auch Trumpf verstärkt den Fokus darauf, seine Geschäftsprozesse und Produkte nachhaltiger zu gestalten. Um die Nachhaltigkeit eines Produktes zu quantifizieren wird eine Kennzahl benötigt, die ausdrückt wie viele Emissionen bei der Herstellung eines Produktes emittiert wurden. Der Product Carbon Footprint (PCF) ist eine Kennzahl, die den Ausstoß verbrauchter Emissionen darstellt. Allerdings existiert bislang keine einheitliche Methode zur genauen Berechnung des PCF. Die Vorgehensweise bei der PCF-Ermittlung variiert stark zwischen verschiedenen Industrien. Die Berechnung des PCF bietet verschiedene Vorteile. Sie zeigt nicht nur auf, wie viele Emissionen tatsächlich anfallen, wenn ein Produkt entsteht, sondern hilft auch dabei, Emissionsquellen entlang des gesamten Lebenszyklus eines Produktes systematisch zu erfassen. Dadurch lassen sich gezielt Maßnahmen entwickeln, um die Effizienz einzelner Produktionsschritte zu verbessern und die Umweltbelastung zu minimieren. In Zukunft wird es unverzichtbar, den PCF auszuweisen, da dies durch regulatorische Vorgaben zunehmend verpflichtend wird. Ein Beispiel dafür ist die Corporate Sustainability Reporting Directive (CSRD) [2], eine EU-Richtlinie, die Unternehmen dazu verpflichtet, umfassend über ihre Nachhaltigkeitsleistung, einschließlich ihrer CO2-Emissionen, zu berichten.

Zielsetzung

Das Ziel dieser Arbeit ist die Entwicklung einer Softwarelösung, die den PCF einer einzelnen Blechkomponente vollautomatisch auf Basis einer Geometriedatei berechnet. Der Schwerpunkt liegt auf der Konzeption und Umsetzung einer Full-Stack-Anwendung, die im internen Trumpf-Netzwerk verfügbar gemacht werden soll. Die Anwendung soll präzise PCF-Werte liefern, um diese im Entwicklungsumfeld zu nutzen. Dadurch sollen fundierte Designentscheidungen ermöglicht werden, die zu einer effizienteren Gestaltung der Blechkomponenten beitragen.

Vorgehensweise

Diese Arbeit orientiert sich an der Design Research Methodology (DRM) 1 nach Blessing und Chakrabarti. Nach einer einführenden Analyse der Fachliteratur erfolgte ein Austausch mit Experten, um die verfügbaren Daten sowie die Anforderungen an eine solche Software zu definieren.

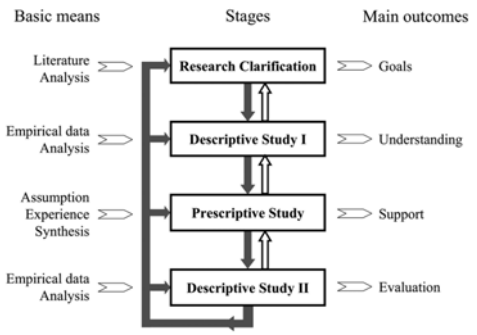


Abb. 1: Phasen der Design Research Methodology [4]

Daraufhin begann die Implementierung, wobei in wöchentlichen Intervallen Feedback zum Fortschritt eingeholt wurde. Dieser iterative Prozess durchlief regelmäßig die drei Phasen der DRM, bis schließlich ein Minimal Viable Product (MVP) erreicht wurde.

Ergebnis

Die Software ist vollständig entwickelt und wird noch in diesem Jahr im Trumpf-Netzwerk bereitgestellt. Ihre Funktionsweise wird im Folgenden beschrieben. Über die zugehörige Webseite (Homepage) 2 können Benutzer Geometriedateien hochladen und dabei Parameter wie Material, Schneidgas, Materialdicke und Lasertyp angeben. Im Hintergrund werden zwei voneinander unabhängige Prozesse ausgeführt. Einer dieser Prozesse berechnet den PCF, während der andere ein Bild basierend auf der Geometriedatei erstellt. Im Rahmen der PCF-Berechnung wird die hochgeladene Datei zunächst aufbereitet und an verschiedene Schnittstellen des Trumpf-Netzwerks übermittelt. Diese Schnittstellen liefern Verbrauchsdaten, die spezifisch für die hochgeladene Geometrie sind. Dazu zählen der Energieverbrauch, die Menge des verwendeten Schneidgases und das eingesetzte Material. Aus diesen Informationen wird der PCF der Geometrie berechnet. Parallel dazu analysiert die Software die Geometriedatei, um Konturen, Kantenlängen und weitere relevante Details zu extrahieren. Diese Daten werden verwendet, um ein Bild der Geometrie zu erstellen. Sobald beide Prozesse abgeschlossen sind, werden im Frontend (Geo-Pcf-Report) 2 sowohl das generierte Bild der Geometrie als auch der berechnete PCF zusammen mit weiteren wichtigen Informationen angezeigt.

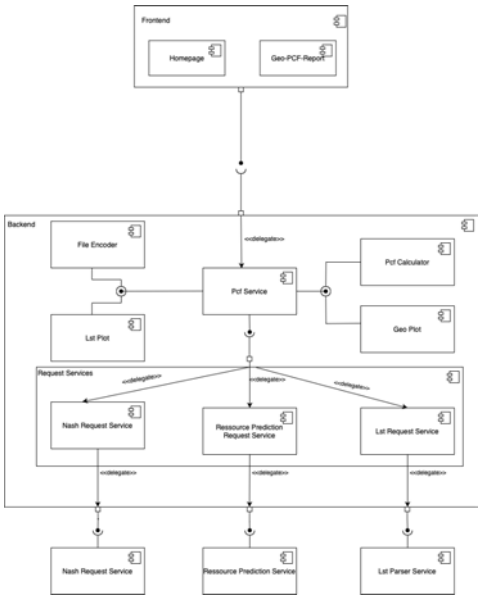


Abb. 2: Komponentendiagramm [1]

Ausblick

Die Anwendung läuft derzeit stabil im Trumpf-Netzwerk und bietet eine verlässliche Basis für zukünftige Erweiterungen. Geplante Weiterentwicklungen umfassen unter anderem die Unterstützung zusätzlicher Dateiformate sowie die Integration einer Kostensicht in die PCF-Berechnung. Dadurch sollen neben den ermittelten CO2-Werten auch entsprechende Kosten ausgewiesen werden, um eine wirtschaftliche Analyse des Carbon Footprints zu ermöglichen. Darüber hinaus ist vorgesehen, die Daten im Frontend durch anschauliche Diagramme zu visualisieren. Dies erleichtert es beispielsweise, schnell zu erkennen, welches verbrauchte Material den größten Einfluss auf den PCF hat.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Susanne Dr Pankov, Eva Kleemann, Elisabeth Voigt, Johanna Anna Hansjürgens, and Alina Ulmer. *EIN KURZER ÜBERBLICK ÜBER DIE EU RICHTLINIE – CORPORATE SUSTAINABILITY REPORTING DIRECTIVE [CSRD]*. adelphi, Thomas Fleissner (DFGE GmbH), 2023.

[3] Simon Dr Schreck, Georg Dr Kobiela, and Simon Dr Wolf. Rahmenbedingungen für die Transformation in Deutschland. *Greenwich*, page 20, 2023.

[4] Sonia Ben Hamida. Innovation by Designing for Value - Towards a Designt-to-Value Methodology in Early Design Stages. https://www.researchgate.net/figure/Design-Research-Methodology-Framework-Blessing-and-Chakrabarti-2009_fig4_325817349, 2017.

Exploration and Evaluation of Multi Camera Asynchronous Fusion for Self-Supervised Monocular Visual Odometry

Jens Wolter **Markus Enzweiler**

Department of Computer Science and Engineering, Esslingen University

Work carried out at Department of Computer Science and Engineering, Esslingen

Introduction

Accurate visual odometry (VO) is essential for autonomous navigation. This is especially true for mobile robots where hardware capabilities are limited. In contrast to autonomous vehicles, these smaller robots lack the capacity to execute complex algorithms such as SLAM due to their limited computational capabilities. In addition to powerful GPUs, these vehicles often use expensive cameras and lidar to further enhance the capabilities of these models.
Our research focuses on these mobile robots with a new and challenging dataset recorded with three basic cameras. In addition, this dataset was recorded in an agricultural environment. This makes it more challenging since a lot of research focuses on more controlled environments, such as indoor or urban driving scenarios, where camera movement is smoother. Specifically, we evaluated a fusion framework that uses self-supervised learning and SCDepth as the pose prediction model.
Our goals are: 1.) evaluate the accuracy of SCDepth as a pose prediction model, 2.) test transfer learning on this model, 3.) explore if enhancements can be made to the SCDepth model to improve its accuracy, 4.) evaluate different fusion modules, and 5.) determine the need for asynchronous fusion and the use of AFT-VO [3].

Multi Camera Monocular VO

VO, describes the task of predicting a traversed path using only the available visual input. Specifically, monocular VO uses only one camera as an input source. This leads to scale ambiguity, where the model is unable to reliably determine how fast the actor is moving. Additionally, VO is susceptible to drift error, wherein the predicted trajectory gradually deviates from the ground truth due to inherent limitations in the estimation process. [5]

A trajectory consists of multiple pose predictions added together. A Pose is calculated from two consecutive images and can be represented as:

$$P_{k,k-1} = [R_{k,k-1}, t_{k,k-1}] \tag{1}$$

where R represents the orientation of the actor and t the translation vector. [5]
Using multiple predicted poses and fusing them together can theoretically improve the accuracy of the predicted pose. If the captured images are asynchronous, a fusion may need to account for the time difference between the different images.
We also employ self-supervised learning, which further complicates this task.

System Architecture

Our desired fusion architecture consists of three design goals: Modularity, Scalability, Efficiency. These design goals allow for the use of different datasets with different numbers of cameras. It also allows the pose prediction or fusion model to be replaced with a better/more accurate model.

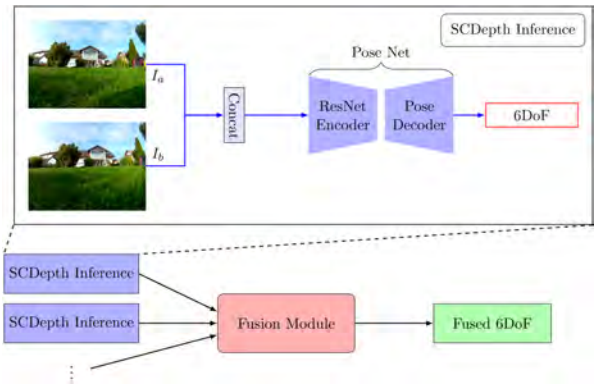


Fig. 1: Inference process of the Fusion Framework, illustrating individual SCDepth inferences and their integration through a fusion module. [4]

Therefore, we decided to base our research on [1]. It introduces a fusion framework; however, its full capabilities could not be thoroughly demonstrated due to time constraints. The fusion framework uses a late fusion strategy, which can be seen in 1.
The pose prediction model used is SCDepth [6]. This model was originally designed for depth estimation and was re-purposed by [1] to predict pose changes. The model is called for each camera, taking two consecutive images as input and returning a pose as output, which are subsequently combined in a fusion module. The three available fusion methods are Naive Fusion, Improved Naive Fusion, and MLP-RNN Fusion. However, our paper does not utilize the Improved Naive Fusion.

Datasets

With our goal to use this fusion model on the mobile robot dataset, we also analyzed three different datasets to see if transfer learning from these could be employed. We decided to additionally only use **DDAD** [2], which was already in use before. However, we made a few changes, the most significant change was that we fixed the orientation in the dataset to follow the track rather then being static all the time. Here we noticed that the Dataset with only 5-10 seconds of data per sequence is to short to be used for reliable comparission between two VO models.
The **mobile robot** was implemented with the same structure as the DDAD Dataset, however we made a few adjustements: 1.) Changed the zero shot depth estimation model from LeReS to a more accurate Depth Anything V2 model. 2.) The trajectories were each divided into three sequences: perimeter, lane, random. 3.) The images were downsampled from 30 Hz to 10 Hz, and the total number of images was standardized across all sequences. 4.) Lastly, this dataset lacks orientational data, which we supplemented using an algorithm. An example can be found in 2.

Results

We found that the DDAD dataset with only 5-10 seconds of data per sequence, was too short to be used as a reliable comparison between two VO models. For asynchronous fusion, we showed that using the AFT-VO approach was not necessary and would not work due to the minimal difference in timestamps between the available images.
The following results where compared, using evo and calculating the RMSE for APE and RPE. They describe the relationship between the estimated trajectory and the ground truth trajectory: APE measures the overall alignment accuracy, while RPE captures the consistency of relative pose estimation between consecutive frames.

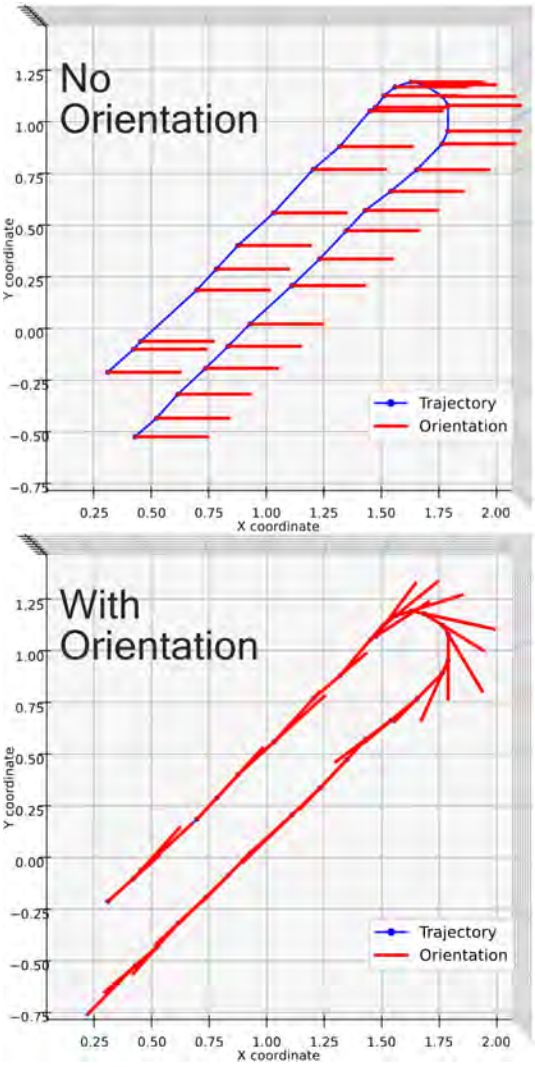


Fig. 2: Comparison of trajectory and orientation before and after applying our method, with a focus on a specific turn in the mobile robot dataset. [4]

After reviewing the predicted trajectories, we found that SCDepth had poor performance for all sequences in the mobile robot dataset. The transfer learning we used also didn't significantly improve the accuracy. We also tried to improve the accuracy of the SCDepth by using a larger model (ResNet-50), not preloading a ResNet, or replacing the ResNet encoder with MobileVitV2. None of these changes improved the accuracy. While we saw a lot of potential for the MLP-RNN fusion, the underlying pose prediction from the SCDepth was too bad to reliably tell if it would have made any significant improvements.

References and figures

[1] Tobias Brandl. Implementation of a Multi-Camera Visual Odometry Fusion Framework to support mobile robot navigation in agricultural environments, 2024.

[2] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3D Packing for Self-Supervised Monocular Depth Estimation. https://arxiv.org/abs/1905.02693, 05 2019.

[3] Nimet Kaygusuz, Oscar Mendez, and Richard Bowden. AFT-VO: Asynchronous Fusion Transformers for Multi-View Visual Odometry Estimation. https://arxiv.org/abs/2206.12946, 2022.

[4] Own representation.

[5] Davide Scaramuzza and Friedrich Fraundorfer. *Visual Odometry [Tutorial]*, volume 18. IEEE Robotics & Automation Magazine, 4 edition, 2011.

[6] Libo Sun, Jia-Wang Bian, Huangying Zhan, Wei Yin, Ian Reid, and Chunhua Shen. SC-DepthV3: Robust Self-supervised Monocular Depth Estimation for Dynamic Scenes. https://arxiv.org/abs/2211.03660, 2022.

Entwicklung und Analyse einer parallelen FPGA-basierten Simulated-Annealing-Architektur für kombinatorische Optimierungsprobleme

Fabian Zaiser **Reiner Marchthaler**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Thales Deutschland, Ditzingen

Field Programmable Gate Arrays (FPGAs) sind heutzutage überall in unserem Alltag. Sie sind in diversen Elektrokleingeräten und Haushaltsgeräten (vor allem Internet of Things) verbaut. Aber auch in weniger alltäglichen Bereichen wie Verteidigung werden FPGA verwendet. Beim Entwerfen von FPGAs drei Schritte essenziell: die Erstellung des Floorplans, die Platzierung von Komponenten und die Verdrahtung dieser Komponenten [4]. Sowohl die Platzierung als auch die Verdrahtung dieser Komponenten stellen beide die Entwickler vor Optimierungsprobleme. In der Praxis werden Algorithmen wie Simulated Annealing verwendet, die das sogenannte Travelling Salesman Problem (TSP) lösen. Sowohl beim TSP als auch beim Routing von Komponenten auf dem FPGA besteht die Herausforderung, wie man die Komponenten (Punkte) am effizientesten verbindet.

Momentan wird das Routing der FPGAs über Simulated Annealing Software seitig optimiert. Dies ist jedoch in der Praxis äußerst rechenaufwändig und man möchte bei Thales evaluieren, ob dieser Algorithmus auch über ein FPGA laufen kann, da Schleifen basierte Algorithmen dort effizienter laufen. Bei der jetzigen Softwareimplementierung besteht ein zeitliches Bottleneck durch das Chipdesign-Implementierungstool nextpnr. Eine Optimierung des SA durch Parallelisierung in nextpnr hat einen erhöhten Stromverbrauch zur Folge, welcher unerwünscht ist. Für die parallelisierte Variante des SA Algorithmuses wird die Übertragung der Daten auf das FPGA in Form von Distanzmatrizen als Bottleneck eingeschätzt, und es stellt sich die Frage, ab welchen Datengrößen die Übertragung der Distanzmatrizen auf das FPGA zu aufwendig wird und der Vorteil durch die effizientere Berechnung auf dem FPGA nicht mehr überwiegt.

Simulated Annealing

Der Simulated Annealing Algorithmus wurde in den 50er Jahren von verschiedensten Mathematikern und Forschern entwickelt. Er beruht auf dem physikalischen Prinzip des Temperns von Metall aus der Thermodynamik. Durch das Abkühlen des heißen Metalls nimmt die kinetische Energie der Teilchen immer weiter ab [1]. Ziel ist es, ein energetisches Minimum zu finden. Der Algorithmus startet zunächst mit einer beliebigen Lösung. Mit dieser Anfangslösung berechnet der Algorithmus eine sogenannte Nachbarlösung. Dies geschieht, indem man einfach 2 Knoten der Anfangslösung vertauscht. Danach werden die Kosten für diese Nachbarslösung berechnet. Im Fall des TSP werden die Kantenlängen zwischen den Punkten nacheinander in der Reihenfolge der Lösungsroute aufsummiert. Sind die Kosten der neuen Lösung berechnet, werden die Kosten der neuen Lösung mit denen der alten verglichen. Sind die neuen Kosten niedriger, wird die neue Lösung als Anfangslösung für die nächste Iteration übernommen. Sind die neuen Kosten höher, wird die Lösung verworfen und die nächste Iteration rechnet mit der gleichen Anfangslösung weiter. Solang, bis eine neue Lösung mit geringeren Kosten gefunden wird oder eine andere Abbruchbedingung erfüllt wird. Das könnte zum Beispiel das Erreichen einer bestimmten Temperatur oder eine zuvor festgelegte Anzahl an Iterationen sein. Wenn der Algorithmus nur bessere Lösungen annimmt, kann es passieren, dass der Algorithmus in einem lokalen Minimum landet. Dieses lokale Minimum wird auch als metastabiler Zustand eines Systems bezeichnet [5]. Das kommt daher, dass die Teilchen bei zu schnellem Abkühlen nicht die Zeit haben, sich richtig zu ordnen. Das hat dann zur Folge, dass das Material instabil ist. Man bezeichnet diesen Zustand auch als gläßern [5]. Um aus diesen lokalen Minima herauszukommen, wendet man beim SA Algorithmus das Prinzip des sogenannten Uphill Climb an. Das Prinzip soll anhand der Abb. 1 erläutert

werden. Wie man an Abb. 1 sieht, gibt es mehrere lokale Minima und Maxima. Um nun aus dem lokalen Maxima zu kommen, akzeptiert der Algorithmus mit einer Wahrscheinlichkeit $p_{accept} = \exp\left(-\frac{\Delta E}{k_B T}\right)$ das Ergebnis trotz schlechterer Kosten. Das ermöglicht dem Algorithmus, von der leicht erhöhten Position aus neue, energetisch optimalere Lösungen zu finden. Mit wieterem Fortschreiten und immer weiter sinkenden Temperaturen, nimmt auch die Wahrscheinlichkeit ab, bei der schlechtere Lösungen noch angenommen werden.

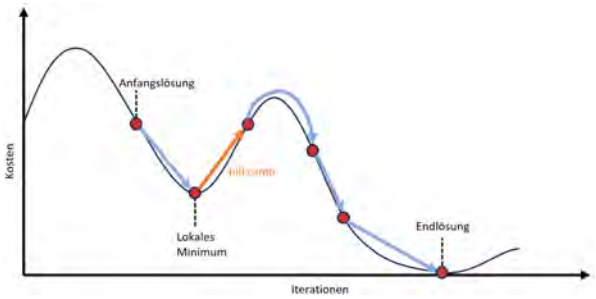


Abb. 1: Optimierungsprozesses mit Hill Climbing [3]

Parallelisierung

Optimierungsalgorithmen können in drei Parallelisierungstypen unterteilt werden [2].**Typ 1** umfasst die parallele Untersuchung innerhalb einer Iteration, wobei alle Prozesse denselben Lösungsraum erkunden, dabei jedoch unterschiedliche Nachbarlösungen prüfen. In Bezug auf die Ausführung der Schritte eines Moves bestehen zwei Varianten: Beim Single-Trial-Ansatz erfolgt eine Aufteilung der Schritte, was zu einer Reduktion der Berechnungszeit pro Move führt. Allerdings ist dies auf einen Move pro Iteration beschränkt. Im Multiple-Trial-Ansatz erfolgt die Berechnung der Moves inklusive der Kostenfunktion durch jeden Prozess eigenständig, wobei die Ergebnisse am Ende einer Iteration miteinander verglichen werden. Im Rahmen von **Typ 2** erfolgt eine Aufteilung des Lösungsraums in disjunkte Teilmengen, welche durch die jeweiligen Prozesse unabhängig voneinander durchsucht werden. Dabei findet das Master-Slave-Prinzip Anwendung. In diesem Zusammenhang ist zu erwähnen, dass bei einem verteilten Speicher Fehler auftreten können, wenn Prozesse mit veralteten Daten arbeiten. Systeme mit gemeinsamem Speicher weisen eine geringere Anfälligkeit auf, da die globale Lösung in kürzeren Abständen aktualisiert wird. Im Verlauf eines Simulated Annealing (SA)-Verfahrens sinkt die Akzeptanzrate mit abnehmender Temperatur, wodurch die Gefahr fehlerhafter Berechnungen durch veraltete Daten reduziert wird.Bei **Typ 3** durchsucht jeder Prozess unabhängig den gesamten Lösungsraum mit unterschiedlichen Heuristiken. Die Ergebnisse werden erst nach Abschluss der Berechnungen miteinander verglichen.

Fazit

Bisher konnte eine einfache, nicht parallelisierte Version des SA Algorithmuses in HLS geschrieben und in Kombination mit Vivado und Vitis IDE auf einem Zynq FPGA zum Laufen gebracht werden. Als Vergleich wurde der Algorithmus in C++ auf einem PC implementiert. Beide Implementierungen wurden mit der gleichen 100x100 Distanzmatrix geladen. Dabei kam es zu folgendem Ergebnis: Die C++ Implementierung erzielte ein von den Kosten her besseres Ergebnis. Die C++ Implementierung erzielte hierbei ein Ergebnis von 2906 als Kosten im Durchschnitt und einer Rechenzeit von durchschnittlich $2,3 * 10^{-3} s$. Die Implementierung kam auf durchschnittliche Kosten von 4556 bei einer Rechenzeit von $4,29 * 10^{-6} s$. Die FPGA Implementierung ist ohne Parallelisierung schon um knapp 500 Mal so schnell, jedoch auch deutlich schlechter, was die Kosten angeht. Dies ist vermutlich auf die Qualität der generierten Zufallszahlen auf dem FPGA zurückzuführen und muss nochmal genauer untersucht werden.

Literatur und Abbildungen

[1] Katharina Al-Shamery. Thermodynamik Teil 6 - Maxwell Boltzmann. https://doi.org/10.5446/43644, 2019.

[2] Teodor Gabriel Crainic and Michel Toulouse. Parallel Strategies for Meta-Heuristic. In *Handbook of Metaheuristics*. Springer US, 2003.

[3] Eigene Darstellung.

[4] Frank Kesel. *FPGA Hardware-Entwurf : Schaltungs- und System-Design mit VHDL und C/C++*. De Gruyter Oldenbourg, 4 edition, 2018.

[5] Scott Kirkpatrick, C. Gelatt, and M. Vecchi. Optimization by Simulated Annealing. In *Science*, volume 220, pages 671–80. American Association for the Advancement of Science, 1983.