HOCHSCHULE
ESSLINGEN

Informatik und
Informationstechnik

IT Innovationen

Band 35

Juni 2025

HOCHSCHULE ESSLINGEN

Grußwort der Fakultät

Liebe Leserinnen und Leser,

Es ist faszinierend, wie manche Weisheiten über Jahrhunderte hinweg Bestand haben. Sokrates erkannte bereits vor über 2000 Jahren: „Je mehr ich weiß, desto mehr weiß ich, dass ich nichts weiß." Jahrhunderte später, im 17. Jahrhundert, drückte Isaac Newton es ähnlich aus, indem er unser Wissen mit einem Tropfen und unser Unwissen mit einem Ozean verglich. Und ganz in unserer Zeit, im Jahr 1999, beschrieben die Psychologen Dunning und Kruger, wie wir oft dazu neigen, unsere eigene Kompetenz zu überschätzen, wenn wir noch nicht viel wissen. Je weniger wir wissen, desto mehr glauben wir zu wissen.

Dieses paradoxe Verhältnis zwischen Unwissen und Wissen haben die Autorinnen und Autoren der Artikel in diesem Band der IT-Innovationen allesamt durchlebt: Zu Beginn der Arbeit an einem wissenschaftlichen Thema scheint alles oft einfach, fast zu einfach. Man fragt sich: „Wie soll ich mich denn damit monatelang beschäftigen?". Doch umso mehr man sich mit dem Thema beschäftigt und an die Umsetzung geht, desto mehr werden einem die Feinheiten und Fallstricke klar. Sonderfälle, die man nicht bedacht hatte. Einschränkungen, die unklar waren. Komplexitäten, die plötzlich sichtbar werden. Und zwischendurch steht dann oftmals die Frage im Raum: „Schaffe ich das überhaupt?" Diese Unklarheit und dieses Unwissen ist für Wissenschaftler normal und wohnt allen wissenschaftlichen Anstrengungen inne. Wenn man wüsste, wie's geht, dann wäre es ja schließlich keine Wissenschaft.

Auch die Arbeiten in diesem Band zeigen genau diesen Weg hin zu neuem Wissen – und auf welch unterschiedliche Weise er beschritten werden kann. Manche setzen sich mit der Frage auseinander, wie generative KI in komplexen Systemen wie dem Personalwesen sinnvoll und verantwortungsvoll eingesetzt werden kann, während andere versuchen, Aktienkurse mit Hilfe von KI zu prognostizieren – und dabei an die Grenzen von Datenqualität und Vorhersagbarkeit stoßen. Wieder andere untersuchen, wie sich die visuelle Konsistenz zwischen Figma-Designs und Webanwendungen automatisiert prüfen lässt, oder entwickeln interaktive Lernplattformen, bei denen KI-gestützte NPCs in virtuellen Welten Wissen vermitteln. Allen gemeinsam ist: Die anfängliche Klarheit weicht mit jedem Fortschritt einer neuen Frage – und gerade daraus entsteht wirklich neues Wissen.

Ich darf Sie auf eine Tauchfahrt in den neuen Teil des Wissens einladen, den unsere Absolventinnen und Absolventen Newtons Ozean der Unwissenheit abringen konnten und wünsche viel Freude bei diesem Ausflug.

Viel Freude beim Lesen wünscht Ihnen

Ihr Prof. Dr. Tobias Heer, Dekan

Die Grenzen der generativen KI

Siyar Adirbelli **Dirk Hesse**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

Heutzutage ist die Künstliche Intelligenz (KI) nicht mehr wegzudenken, da sie in verschiedenen Bereichen des Lebens zum Einsatz kommt. Derzeit tendiert die Entwicklung in Richtung einer sogenannten hybriden Intelligenz, die sowohl menschliche als auch künstliche Intelligenz enthält. Dennoch sollte man sich ihrer Schwächen und Grenzen bewusst sein, denn auch die KI stößt immer wieder an ihre Grenzen, sei es in technischer, ethischer oder rechtlicher Sicht. Das Ziel besteht darin, statistische Lernalgorithmen mit logischen und wissensbasierten Methoden zu verbinden. Dieser Ansatz ist vergleichbar mit der Verbindung von unbewusster Wahrnehmungsverarbeitung und bewusstem logischem Schließen im menschlichen Organismus. Hybride KI-Systeme werden mit höheren Intelligenzgraden in Verbindung gebracht [3].

Theoretische Grundlagen

Künstliche Intelligenz basiert aus von Menschen konzipierten Systemen. Hierbei finden sie sowohl in der physischen als auch digitalen Welt ihren Einsatz. Als Grundlage ihres Wissens bezieht sie sich auf ihre Umgebung sowie Daten und leitet daraus das, aus ihrer Sicht, bestmögliche Ergebnis. KI kann zudem, auf Basis früherer Daten, lernen, ihre Aktionen anhand ihrer Umgebung anzupassen [3]. Generative KI ist eine Form der künstlichen Intelligenz, deren Hauptaufgabe das Erstellen von Inhalten ist, wie beispielsweise Text, Audio, Bilder oder Videos. Grundlage der generativen KI ist das sogenannte Machine Learning, das ermöglicht, aus Datensätzen zu lernen. Hierbei wird nicht nur aus den Daten gelernt, sondern auch neue Datensätze auf Basis des Erlernten erzeugt. Noch intensiver arbeitet dabei das Deep Learning, das die Funktionsweise des menschlichen Gehirns annimmt (siehe Abbildung 1). Die generative KI kann zudem in unterschiedliche Arten eingeteilt werden. Zum einen gibt es die sogenannten „Transformator-basierte Modelle“, die besser als GPT-3 oder GPT-4 bekannt sind. Diese sind vor allem für die Erstellung eines Textes hilfreich, da sie die gesamte Eingabe berücksichtigen. Des Weiteren gibt es sogenannte „Generative kontradiktorische Netze“. Diese enthalten einen Generator und einen Diskriminator. Dabei ist der Generator für die Erstellung neuer Dateninstanzen und der Diskriminator für die Überprüfung dieser zuständig. Aufgabe des Generators ist es Daten herzustellen, die der Diskriminator versucht von echten Daten zu unterschieden. Eine andere Art ist der „Variationale Autoencoder“. Dieser codiert wie sein Name schon verrät Eingabedaten, um am Ende dann neue Daten durch die Dekodierung zu erhalten. Durch den Einsatz eines Zufallsfaktors erhält er dann ähnliche Dateninstanzen [4].

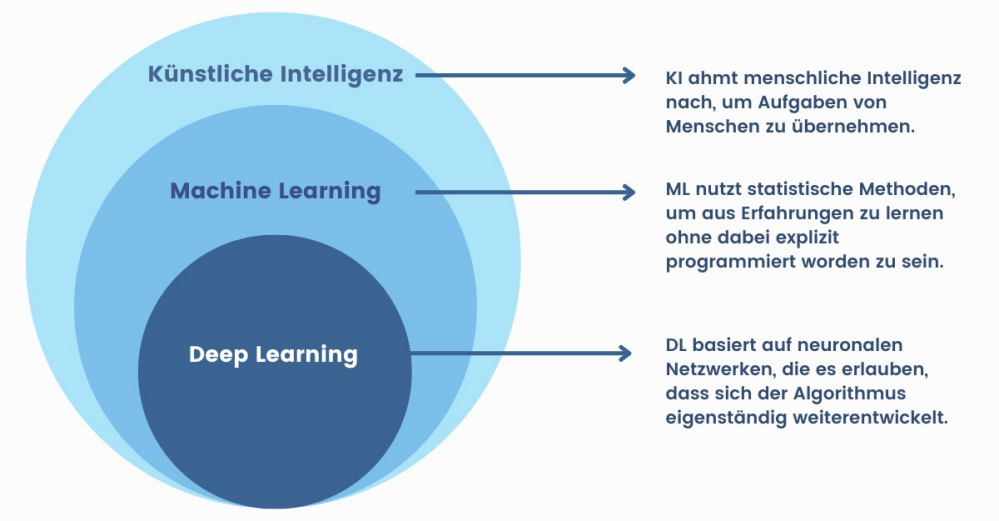


Abb. 1: Grundlage KI [5]

Grenzen der generativen KI

Generative KI hat viele Vorteile und sie hat viele Einsatzmöglichkeiten. Trotzdem gibt es auch einige Punkte, die man nicht außer Acht lassen sollte. Denn die Technologie hat ihre Grenzen. Bei der Auswahl von HR-Tools ist der Datenschutz besonders wichtig. Die Verarbeitung sensibler Informationen durch Plattformen wie ChatGPT bringt ein gewisses Risiko mit sich, insbesondere bei Speicherung der Daten außerhalb der EU. In diesen Ländern gelten oft andere Datenschutzvorschriften, die nicht mit unseren Standards übereinstimmen. Aus diesem Grund empfiehlt es sich, bei der Auswahl auch lokale Alternativen in die Überlegungen mit einzubeziehen. Ein weiterer, oft unterschätzter, Punkt ist die Einbindung in bestehende HR-Systeme. Ein intelligenter Chatbot benötigt nicht nur Wissen, sondern muss auch die jeweilige Situation verstehen. Die Verwendung von Standardantworten ist in der Praxis oft nicht hilfreich. Damit ein Bot wirklich passende Antworten liefern kann, muss er wissen, mit wem er spricht. Die Antwort ist umso genauer, je genauer die Frage oder der Auftrag ist. Solche Informationen sollten idealerweise automatisch aus dem Stammdatensystem kommen. Das betrifft etwa den Arbeitsort oder die Funktion. Generative KI kennt keine firmenspezifischen Prozesse oder interne Richtlinien, was eine klare Grenze ist. Wenn die KI nicht an die individuellen Bedürfnisse angepasst wird, kann sie oft nicht auf die speziellen Fragen aus dem Personalmanagement reagieren. Manchmal ist es nicht möglich, Systeme einfach mit internem Wissen zu erweitern. Aus diesem Grund sollte man bei der Auswahl darauf achten, dass die genutzte Lösung eine einfach steuerbare, sichere und transparente Datenbasis bietet. Des Weiteren lassen sich nicht alle HR-Prozesse ohne Weiteres in eine generative KI überführen. Wenn die Technologie nicht flexibel genug ist, kann das zu ineffizienten Abläufen führen. Das gilt insbesondere für komplexere oder sehr spezielle Aufgaben. Zudem neigt generative KI gelegentlich dazu, Inhalte zu erfinden, ein Phänomen, das als „Halluzination“ bekannt ist. Dieses Problem ist bislang nicht vollständig gelöst. Damit die Informationen nicht falsch sind, muss man immer wieder überprüfen und die Modelle, auf denen sie basieren, immer wieder verbessern [1].

Generative KI kann nur mit dem arbeiten, was ihr bekannt ist, also mit den Daten, auf denen sie basiert. Das heißt also, dass alles, was eine KI ausgibt, direkt davon abhängt, welche Informationen ihr vorher zur Verfügung gestellt wurden. In der Regel stammen diese Daten aus dem Training der KI. Bei einer Lösung, die nicht selber entwickelt wurde, hat man in der Praxis kaum Einfluss auf die Verwendung der Daten für das Training. Das Risiko hierbei ist, dass es möglich ist , dass Teile dieser Trainingsdaten problematisch oder sogar unethisch sind. Beispielsweise wenn sie geschützt sind und ohne Erlaubnis benutzt wurden. Oder wenn sie von schlechten Quellen kommen. Oder wenn sie Vorurteile und Ungerechtigkeiten verbreiten. Besonders schwer wird es, wenn die KI in sensiblen Bereichen eingesetzt wird, etwa im Recruiting. Wenn ein System beispielsweise nur mit Daten gefüttert wurde, die ein sehr einseitiges Bild eines idealen Bewerbers zeichnen, übernimmt es dieses Vorurteil. Im schlimmsten Fall werden dann nur noch junge, blonde Frauen für eine Stelle bevorzugt, weil das in den Trainingsdaten so eingestellt war. Das kann zu diskriminierenden Entscheidungen führen, ohne dass es auf den ersten Blick auffällt [2].

Ausblick

Die Entwicklung generativer KI schreitet schnell voran und eröffnet sowohl in der Forschung als auch in der Praxis neue Möglichkeiten. Die theoretischen Grundlagen, wie Machine Learning, Deep Learning und verschiedene Modellarten (z. B. GPT-Modelle, GANs oder Autoencoder), verdeutlichen die Leistungsfähigkeit moderner KI-Systeme. Sie analysieren große Datenmengen, identifizieren Muster und generieren neue Inhalte, die in zahlreichen Anwendungsbereichen einen echten Vorteil bieten.
Im praktischen Einsatz zeigt sich jedoch, dass diese Technologie gewisse Herausforderungen mit sich bringt. Es wird deutlich, dass die Qualität und Herkunft der Trainingsdaten einen entscheidenden Einfluss auf die Leistung der generativen KI hat. An dieser Stelle ergeben sich Risiken, wie beispielweise das Fehlen an Transparenz und Kontrolle über diese Daten kann dazu führen, dass unethische Inhalte, Urheberrechtsverletzungen oder diskriminierende Strukturen unbemerkt übernommen und wiederverwendet werden. Insbesondere in sensiblen Bereichen wie dem Personalwesen oder Recruiting können die Konsequenzen erheblich sein, sowohl in rechtlicher als auch in imagebezogener Hinsicht.
Des Weiteren wird deutlich, dass generative KI-Systeme in ihrer gegenwärtigen Form Schwierigkeiten haben, firmenspezifisches Wissen oder unternehmensinterne Prozesse zu berücksichtigen, ohne dass dies mit einem hohen Aufwand verbunden wäre. Ohne eine gezielte Anpassung ist die KI auf allgemeine Aussagen beschränkt, die im praktischen Arbeitsalltag oft nicht ausreichen. Auch die technische Integration in bestehende Systeme, wie etwa HR-Plattformen, stellt Unternehmen vor strukturelle Herausforderungen. Die Ausstattung eines Chatbots mit Basiswissen ist nicht ausreichend. Vielmehr sind Schnittstellen, saubere Datenstrukturen sowie die Fähigkeit, Kontextinformationen wie Rolle, Standort oder Funktionsstufe zu erkennen und sinnvoll zu verarbeiten, erforderlich.

Literatur und Abbildungen

[1] Felix Anderegg and Esther Brand. Grenzen der generativen KI: Stolpersteine und Grenzen von ChatGPT im HR. https://www.weka.ch/themen/personal/personalfuehrung-und-personalentwicklung/digitalisierung-im-hr/article/grenzen-der-generativen-ki-stolpersteine-und-grenzen-von-chatgpt-im-hr/, 2024.

[2] Maren Dinges. Ethik der generativen KI – Grenzen und Verantwortlichkeiten beim Einsatz von KI zur Inhaltserstellung. https://simpleshow.com/de/blog/ethik-generativen-ki-grenzen-verantwortlichkeiten-einsatz-ki-inhalteerstellung/, 2024.

[3] Klaus Mainzer and Reinhard Kahle. *Grenzen der KI theoretisch, praktisch, ethisch*. Springer-Verlag GmbH, 2022.

[4] Firma SAP. Was ist generative KI. https://www.sap.com/germany/products/artificial-intelligence/what-is-generative-ai.html, 2024.

[5] Firma Stackfuel. Was ist Machine Learning? Algorithmen, Methoden und Beispiele 2024. https://stackfuel.com/de/blog/machine-learning-algorithmen-data-analytics/, 2024.

Konzeptionierung und Ausarbeitung eines automatisierten Datenmanagementsystems für eine transparente Verfolgung effizienzsteigernder Maßnahmen in den administrativen Bereichen der MAHLE GmbH

Nick Albrecht **Markus Enzweiler**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma MAHLE GmbH, Feuerbach

Einleitung

Lean Management, das aus den Ansätzen des Toyota Production System (TPS) hervorgegangen ist, hat seit den neunziger Jahren stark an Bedeutung gewonnen [6]. Dabei handelt es sich nicht um eine einzelne Methode, sondern um einen kontinuierlichen Ansatz zur Prozessoptimierung. Im Mittelpunkt stehen dabei Kundenorientierung und Kostenreduktion [1]. Durch die zunehmende Globalisierung und weltweite Vernetzung entsteht ein hoher Kosten- und Wettbewerbsdruck [1]. Ein effektives Lean Management trägt dazu bei, wettbewerbsfähig zu bleiben und gleichzeitig Flexibilität in Bereichen wie Verfügbarkeit, Qualität und Service zu gewährleisten [1], [7].

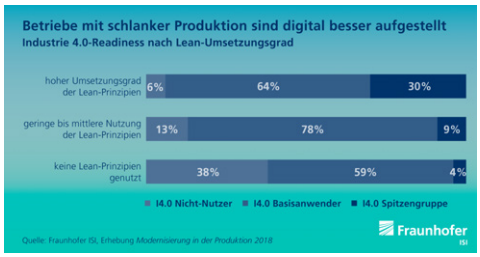


Abb. 1: Industrie 4.0 Readiness nach Lean Umsetzungsgrad [7]

Auch bei der MAHLE GmbH wird dieser Lean-Ansatz verfolgt. Zu diesem Zweck wurde die Initiative E:MC² (Efficiency: Motivation & Continuous Improvement & Culture) ins Leben gerufen. Mit verschiedenen Methoden wird der Lean-Gedanke gezielt in die administrativen Bereiche übertragen. Um Fortschritte sowie Zeit- und Kosteneinsparungen zu dokumentieren, kommen verschiedene Tools zum Einsatz, mit denen **Daten** aus unterschiedlichen Abteilungen erfasst werden.

Ziel der Arbeit

Das Ziel dieser Bachelorarbeit ist es, ein Datenmanagementsystem zu entwickeln, das Daten aus verschiedenen Quellen miteinander verknüpft, analysiert, visualisiert und zur Erstellung von Prognosen sowie zur Erkennung von Tendenzen genutzt werden kann. Aktuell werden die benötigten Daten größtenteils manuell in Excel-Tabellen zusammengetragen. Das ist nicht nur zeitaufwändig und fehleranfällig, sondern erschwert zudem die Weiterverarbeitung der Daten, insbesondere im Hinblick auf den Einsatz moderner, KI-gestützer Analysemethoden. In diesem Artikel wird zunächst ein Überblick über das Thema Data Science und das Microsoft-Programm Power BI gegeben. Anschließend werden das Vorgehen und der weitere Ausblick beschrieben.

Data Science

Data Science beschäftigt sich mit der Gewinnung von Wissen aus Daten. Dabei werden aus großen Datenmengen relevante Informationen extrahiert, um daraus Handlungsempfehlungen für das Management und das Unternehmen abzuleiten. Ziel dieser Empfehlungen ist es, die Qualität unternehmerischer Entscheidungen zu verbessern und die Effizienz von Arbeitsprozessen zu steigern. Data Science vereint Elemente aus Mathematik, Statistik, Informatik sowie branchenspezifischem Fachwissen, um Anomalien in Daten zu erkennen und Vorhersagen über zukünftige Ereignisse treffen zu können [9]. Im Wesentlichen wird der Data-Science-Lebenszyklus in fünf Hauptphasen unterteilt (siehe Abb. 2). Die ersten beiden Phasen befassen sich mit der Datensammlung und -aufbereitung. In der dritten und vierten Phase werden die Daten analysiert und visualisiert, wobei häufig Machine Learning Algorithmen zum Einsatz kommen. Die letzte Phase dient der Interpre-

tation und Kommunikation der aus der Datenanalyse gewonnenen Ergebnisse [8]. Auf Basis dieser Erkenntnisse und Prognosen können anschließend fundierte Entscheidungen getroffen werden, die beispielsweise der Kosten- oder Umsatzoptimierung dienen. Dabei unterstützen Machine-Learning-Algorithmen zunehmend automatisierte Entscheidungsprozesse auf Grundlage der vorhandenen Daten. Durch diesen Lebenszyklus ist es möglich, datenspezifische Projekte systematischer und erfolgreicher umzusetzen [9].

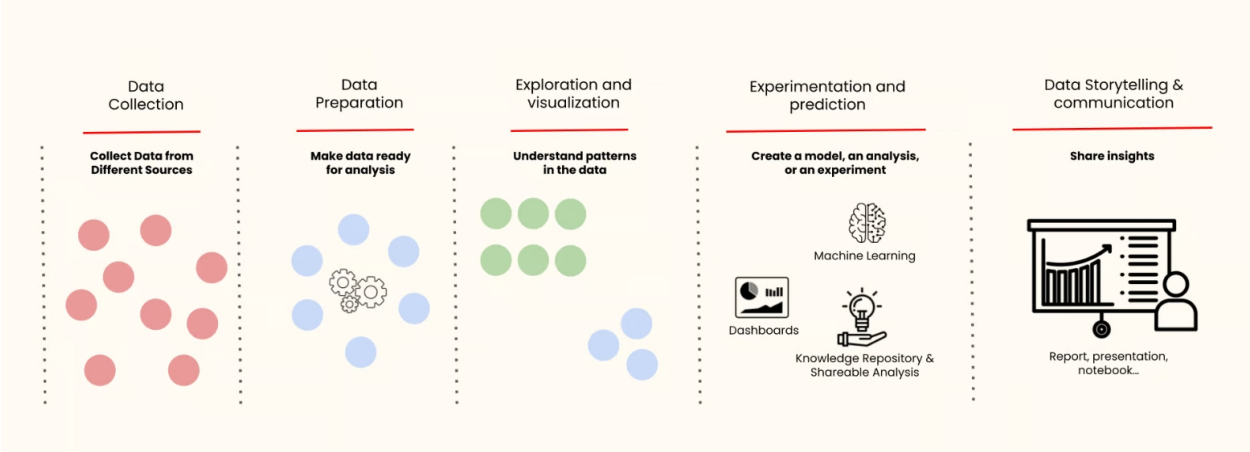


Abb. 2: Der Lebenszyklus von Data Science [8]

Power BI

„Power BI ist eine Sammlung von Softwarediensten, Apps und Connectors, die zusammenarbeiten, um ihre nicht verbundenen Datenquellen in kohärente, visuell überzeugende und interaktive Erkenntnisse umzuwandeln"[3]. Dabei können Daten in unterschiedlichen Formen, wie zum Beispiel einer Excel Tabelle, aus einem SAP System oder aus anderen cloudbasierten oder lokalen Data Warehous-Instanzen vorliegen. Mithilfe eines Datenmodells (siehe Abb.3) können Verbindungen zwischen den verschiedenen Datenquellen hergestellt werden. Power BI ermöglicht es außerdem die Daten aufzuarbeiten, zu analysieren und zu visualisieren. Dabei können die Berichte oder Dashboards mit mehreren Endgeräten geteilt werden [3].

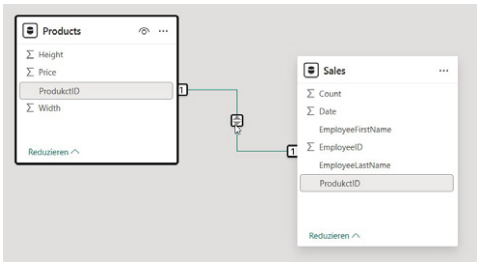


Abb. 3: Power BI-Datenmodell - Beispiel für die Verknüpfung zweier Dateien über den gemeinsamen Eintrag ProductID. [2]

Vorgehen

Um die unterschiedlichen Daten in ein geeignetes System zu überführen, werden sie zunächst gesammelt und strukturiert. Anschließend erfolgt die Übertragung in Power BI, wo sie auf relevante Aspekte reduziert und im Datenmodell miteinander verknüpft werden. Diese Bereinigung und Aufbereitung der Daten erfolgt mithilfe der in Power BI integrierten Programmiersprache M (Power Query-Formelsprache) [5]. Für die Modellierung und Verknüpfung der Daten im Datenmodell werden Data Analysis Expressions (DAX) verwendet [4]. Sobald die Daten aufbereitet sind, können sie im nächsten Schritt visualisiert und analysiert werden. In einer geschlossenen Workspace-Umgebung lassen sich Reports und Dashboards erstellen, auf denen wesentliche Informationen übersichtlich dargestellt werden. Dazu zählen insbesondere bisher erzielte Zeit- und Kosteneinsparungen in verschiedenen Bereichen sowie weitere Kennzahlen, die einzelne Abteilungen im Vergleich zum Gesamtkontext des Unternehmens einordnen. So kann beispielsweise analysiert werden, welche Bereiche den Lean-Ansatz besonders erfolgreich umgesetzt haben und es lassen sich Hypothesen über die Ursachen ableiten und welche Maßnahmen gegebenenfalls notwendig sind, um andere Bereiche zu verbessern. Auch für das E:MC²-Team entsteht auf diese Weise wertvolles Feedback: In welchen Bereichen kamen die Inhalte möglicherweise nicht wie gewünscht an? Bei welchen Führungskräften besteht noch Unterstützungsbedarf? Darüber hinaus können die Reports und Dashboards auch Tendenzen und Prognosen enthalten, also Annahmen darüber, wie

bestimmte Kennzahlen oder Entwicklungen künftig aussehen könnten.

Ausblick

Die durch das Datenmanagementsystem gewonnenen Informationen können genutzt werden, um Lean Management und damit die Effizienzsteigerung von internen Geschäftsprozessen bei MAHLE weltweit weiter voranzutreiben. Gleichzeitig bietet auch das System selbst Potenzial zur Weiterentwicklung: Eine optimierte Datensammlung kann beispielsweise qualitativ hochwertigere Daten liefern, die noch bessere Analysen ermöglichen. Ziel ist es, ein Datenmanagementsystem zu entwickeln, das sich kontinuierlich verbessern lässt, ganz im Sinne des Continuous Improvement nach dem E:MC^2 Prinzip.

Literatur und Abbildungen

[1] Christian Böning. Lean Management. https://www.iph-hannover.de/de/information/lean-production/lean-management/#:~:text=Das%20Lean%20Management%20stellt%20keine,kompletten%20Wer, 2025.

[2] Eigene Darstellung.

[3] David Iseminger et al. Was ist Power BI? https://learn.microsoft.com/de-de/power-bi/fundamentals/power-bi-overview, 03 2024.

[4] David Iseminger et al. Erlernen der DAX-Grundlagen in Power BI Desktop. https://learn.microsoft.com/de-de/power-bi/transform-model/desktop-quickstart-learn-dax-basics, 04 2025.

[5] Doug Klopfenstein and Jason Howell. Power Query M – Formelsprache. https://learn.microsoft.com/de-de/powerquery-m/m-spec-introduction, 07 2023.

[6] Ayelt Komus and Waldemar Kamlowski. Gemeinsamkeiten und Unterschiede von Lean Management und agilen Methoden. https://www.hs-koblenz.de/fileadmin/media/fb_wirtschaftswissenschaften/Forschung_Projekte/Forschungsprojekte/BPM-Labor/BPM-Lab-WP-Lean-vs-Agile-v1.0.pdf, 05 2014.

[7] Christian Lerch. Lean 4.0: Besseres Wirtschaften durch die Kombination smarter und schlanker Produktionsmethoden. https://www.isi.fraunhofer.de/de/presse/2021/presseinfo-21-lean-management.html, 09 2021.

[8] Adel Nehme and Matt Crabtree. Was ist Data Science? Definition, Beispiele, Tools & mehr. https://www.datacamp.com/de/blog/what-is-data-science-the-definitive-guide, 09 2024.

[9] Laurenz Wuttke. Was ist Data Science? Definition, Anwendung und Beispiele. https://datasolut.com/was-ist-data-science/, 02 2024.

Anwendung von Künstliche Intelligenz in der Finanzanalyse, Einsatz von KI zur Vorhersage von Aktienkursen und zur Risikobewertung

Ali Ramin Alizada **Catharina Kriegbaum-Kling**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

Die Aktienanalyse ist ein faszinierendes und komplexes Feld, das von zahlreichen Faktoren beeinflusst wird, die oft schwer vorherzusagen sind. Künstliche Intelligenz (KI) hingegen ist darauf ausgelegt, komplexe Aufgaben zu bewältigen und Muster in großen Datenmengen zu erkennen. Während Experten in der Finanzwelt ihr Bestes tun, um zukünftige Kursentwicklungen vorherzusagen, könnte es sein, dass KI in naher Zukunft diese Aufgabe übernimmt und sie zu einer Routineaufgabe macht. Diese Untersuchung zielt darauf ab, herauszufinden, inwieweit KI tatsächlich in der Lage ist, Aktienkurse zuverlässig vorherzusagen. Durch den Vergleich von KI-gestützten Vorhersagen mit den Einschätzungen von Finanzexperten wird analysiert, ob KI eine geeignete Lösung für die Herausforderungen der Aktienprognose darstellen kann.

Grundlage der Aktienanalyse

Die Aktienanalyse hat das Ziel, den inneren Wert von Aktien zu ermitteln und zukünftige Kursentwicklungen zu prognostizieren. Dabei werden verschiedene Faktoren untersucht, die den Wert eines Unternehmens beeinflussen. Zwei zentrale Methoden stehen im Fokus dieser Analyse: die fundamentale Analyse und die technische Analyse. Beide Ansätze zusammen ermöglichen fundierte Investmententscheidungen. Abbildung 1 veranschaulicht die grundlegende Struktur der Aktienanalyse, die sich in die folgenden zwei Bereiche gliedert:

- Fundamentale Analyse Die fundamentale Analyse bewertet Unternehmen anhand wirtschaftlicher Kennzahlen wie dem Kurs-Gewinn-Verhältnis (KGV), EBITDA und Dividenden. Zudem werden die Branchenlage, gesetzliche Rahmenbedingungen, Steuern und Wettbewerbsbedingungen sowie makroökonomische Daten wie Zinsen, Zölle und Rohstoffpreise untersucht. Mit diesen Bewertungsmodellen versucht die Fundamentalanalyse, den inneren Wert einer Aktie zu bestimmen, von dem aus zukünftige Kursentwicklungen vorhergesagt werden können.
- Technische Analyse Die technische Analyse konzentriert sich auf Kursverläufe und Handelsvolumen, um zukünftige Kursentwicklungen anhand historischer Daten vorherzusagen. Verschiedene Chartarten wie der Kerzenchart, visualisieren Kursverläufe und Trends. Logarithmische Skalen sind für längere Zeiträume mit starken Kursveränderungen geeignet, da sie prozentuale Bewegungen besser darstellen. Technische Indikatoren erweitern die Analyse, indem sie Faktoren wie Preise, Marktstimmung und erwartete Cashflows einbeziehen, die Angebot und Nachfrage beeinflussen. Beispiele für Indikatoren sind der Gleitende Durchschnitt, der RSI (Relative Stärke Index), der MACD (Moving Average Convergence Divergence) und BB (Bollinger Bänder), die zusammen mit Aktienkursen visualisiert werden.

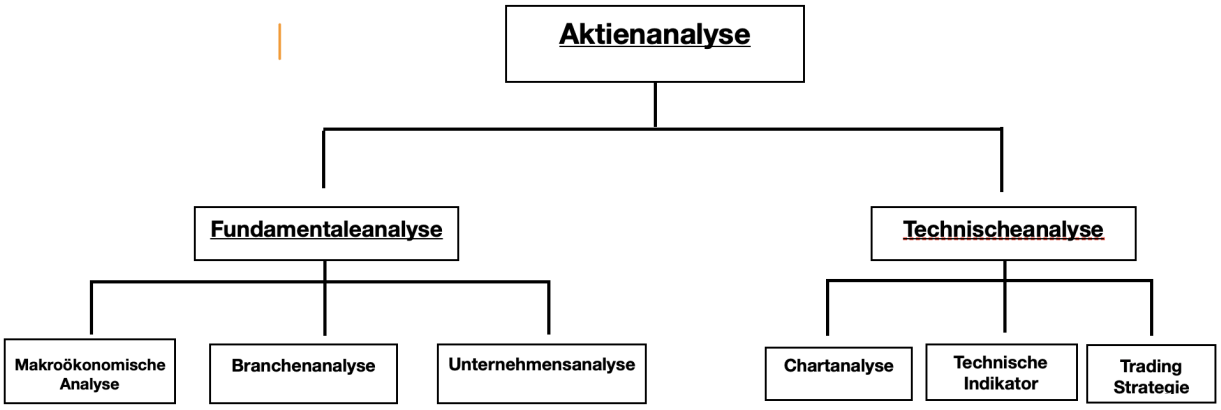


Abb. 1: Aktienanalyse Aufbau [1]

Grundlagen der Künstliche Intelligenz

Künstliche Intelligenz (KI) lässt sich in zwei Hauptkategorien unterteilen: schmale KI und allgemeine KI. Schmale KI, auch schwache KI genannt, ist auf spezifische Aufgaben wie Spracherkennung oder Bilderkennung spezialisiert und wird bereits in Anwendungen wie Sprachassistenten und autonomen Fahrzeugen eingesetzt. Sie basiert auf Technologien wie maschinellem Lernen und neuronalen Netzen. Die symbolische Methode von KI verfolgt einen deduktiven Ansatz, während die subsymbolische Methode auf induktives Lernen setzt.

- Symbolische Künstliche Intelligenz: verwendet einen deduktiven Ansatz mit logischen Regeln zur Wissensrepräsentation, z. B. durch Ontologien, semantische Netze und Wissensgraphen. Für unsicheres Wissen kommen probabilistische Methoden wie die Bayes'sche Inferenz zum Einsatz.
- Subsymbolische Künstliche Intelligenz: basiert auf einem induktiven Ansatz und unterscheidet zwischen überwachten und unüberwachten Lernmethoden. Überwachtes Lernen nutzt vorgegebene Zielparameter, während unüberwachtes Lernen keine vordefinierten Ziele benötigt. Teilüberwachtes Lernen kombiniert beide Ansätze. [3]

Anwendung von KI in der Finanzanalyse

Um die Zusammenhänge einzelner Faktoren und ihre Abhängigkeit voneinander zu untersuchen, wurden in einer Studie mithilfe von ML-Methoden nahezu 1,9 Mrd. Aktien-Monat-Beobachtungen aus dem Zeitraum von 1980 bis 2019 in 68 Ländern ausgewertet. Diese ML-Methoden sind in der Lage „komplexe Beziehungen innerhalb großer Datensätze aufzudecken". Laut dieser Studie hätten die KI-Modelle signifikant besser abgeschnitten als herkömmliche Methoden. „Die maschinellen Lernmodelle können Aktienrenditen mit bemerkenswerter Genauigkeit vorhersagen, sie erzielen monatlich eine durchschnittliche Rendite von bis zu 2,71 % im Vergleich zu herkömmlichen Verfahren mit einer monatlichen Überrendite von etwa einem Prozent." Essentiell für die KI-Modelle ist jedoch wie immer eine saubere Datenbasis, je sauberer die Daten, desto genauer werden die Ergebnisse. [2]

Fallstudie mit Intellecia.ai und Kavout

Das Ziel dieser Untersuchung ist es, mithilfe von Intellecia AI und Kavout die Aktienkurse von SAP, Microsoft und Oracle täglich vorherzusagen und die Ergebnisse über einen Monat zu dokumentieren. Diese Vorhersagen werden mit den tatsächlichen Kursentwicklungen und den Prognosen von Finanzexperten verglichen. Das Ergebnis soll zeigen, ob KI in der Lage ist, Aktienkurse zuverlässig vorherzusagen und welche Stärken und Schwächen sie aufweist. Intellecia.ai (Intellectia AI) ist eine KI-gestützte Investitionsplattform, die Anlegern hilft, informierte Entscheidungen zu treffen. Sie bietet Echtzeit-Marktdaten, technische Analysen und personalisierte Empfehlungen, um die Komplexität finanzieller Analysen zu reduzieren. Zusätzlich wird auch Kavout untersucht, eine weitere Plattform, die maschinelles Lernen zur Bewertung von Aktien verwendet. Kavout hebt sich durch den "K Score"hervor, einen Algorithmus, der auf umfangreichen Datenanalysen basiert und potenzielle Gewinneraktien identifiziert.

Ausblick

Die Vorhersage von Aktienkursen mithilfe Künstlicher Intelligenz ist grundsätzlich möglich, doch eine präzise und verlässliche Prognose bleibt schwierig. Aktienkurse werden nicht nur durch unternehmensspezifische Kennzahlen und technische Daten beeinflusst, sondern auch durch externe, schwer vorhersehbare Faktoren wie politische Entscheidungen, Zinspolitik, internationale Handelsabkommen oder plötzliche geopolitische Ereignisse. Diese Vielzahl an Einflussgrößen macht eine exakte Vorhersage selbst für moderne KI-Systeme herausfordernd.

Literatur und Abbildungen

[1] Fatima Dakalbab, Manar Abu Talib, Qassim Nasir, and Tracy Saroufil. Artificial intelligence techniques in financial trading: A systematic literature review. *Journal of King Saud University - Computer and Information Sciences*, page 3, 2024.

[2] Sylvia Meier and Angelika Breinich-Schilly. *Finance + Banking*. Springer Gabler, 2024.

[3] Wolfgang Wahlster and Christoph Winterhalter. *DEUTSCHE NORMUNGSROADMAP KÜNSTLICHE INTELLIGENZ*. DIN,DKE, 2022.

KI-gestützte agile Methoden: Wie Künstliche Intelligenz das Projektmanagement verändert und unterstützt

Anas Alrzg **Manfred Schoch**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Abstract

Diese Arbeit untersucht, wie Künstliche Intelligenz (KI) agile Methoden im Projektmanagement unterstützen kann. Ziel ist es, typische Herausforderungen wie unklare Aufgabenverteilung oder fehlende Risikofrüherkennung durch KI-gestützte Tools zu verbessern. Der Artikel zeigt auf, welche Potenziale KI dabei bietet und welche Grenzen aktuell noch bestehen.

1. Einleitung

Agile Methoden wie Scrum oder Kanban helfen Unternehmen, Projekte schneller und flexibler umzusetzen. Doch viele Teams stoßen in der Praxis an ihre Grenzen – besonders bei komplexen Projekten. Künstliche Intelligenz (KI) könnte hier unterstützen. Aber wie genau? Und funktioniert das wirklich? Diese Fragen werden in dieser Bachelorarbeit untersucht.

2. Problemstellung

In der Realität zeigt sich: Agile Teams verlieren oft den Überblick über Aufgaben, Abhängigkeiten und Risiken. Entscheidungen basieren häufig auf Bauchgefühl statt auf Daten. Genau hier fehlt es vielen Unternehmen an unterstützenden Tools, die diese Lücken schließen.

3. Ziel der Arbeit

Ziel der Arbeit ist es, herauszufinden, wie Künstliche Intelligenz agile Teams im Projektmanagement unterstützen kann – von der Planung über die Aufgabenverteilung bis zur Risikoanalyse. Dabei geht es auch um die Frage, wie Mensch und Maschine sinnvoll zusammenarbeiten können, ohne die Prinzipien agiler Arbeit zu gefährden.

4. Vorgehensweise

Auf Basis einer systematischen Literaturrecherche wurden zentrale Themenfelder an der Schnittstelle von agilen Methoden und Künstlicher Intelligenz identifiziert. Darauf aufbauend wurde ein strukturierter Fragebogen entwickelt, der gezielt auf typische Herausforderungen, Potenziale und Anforderungen im agilen Projektmanagement eingeht.

Im Rahmen der empirischen Untersuchung ist geplant, vier bis fünf Expertinnen und Experten aus der Praxis zu befragen, die über Erfahrung sowohl im agilen Arbeiten als auch im Umgang mit KI-gestützten Tools verfügen. Die Befragung erfolgt leitfadengestützt. Die einzelnen Antworten werden pro Frage dokumentiert und anschließend systematisch ausgewertet.

Zur Analyse werden die Aussagen je Frage thematisch zusammengefasst und in Form stichpunktartiger Kernaussagen dargestellt. Dabei werden Gemeinsamkeiten, Unterschiede sowie besonders markante Positionen herausgearbeitet. Die strukturierte Darstellung erfolgt in tabellarischer Form.

Ziel ist es, auf Basis der Aussagen zu identifizieren, welche konkreten Probleme und Potenziale im Zusammenspiel von KI und agilen Methoden bestehen und welche Anforderungen sich daraus für den praktischen Einsatz ableiten lassen.

Abb. 1: Forschungsdomänen der Untersuchung [2]

5. Hintergrund und theoretische Basis

5.1. Agile Methoden im Projektmanagement

Agile Methoden wie Scrum oder Kanban helfen Teams, schneller zu arbeiten und sich flexibel an neue Anforderungen anzupassen. Sie setzen auf kurze Entwicklungszyklen, regelmäßiges Feedback und selbstorganisierte Teams. Ziel ist es, Projekte Schritt für Schritt weiterzuentwickeln und früh auf Änderungen zu reagieren [5].

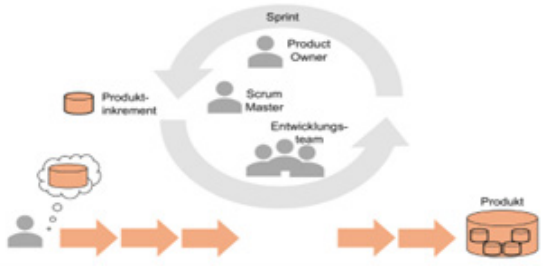


Abb. 2: Scrum Vorgehens weise [3]

Abbildung 2: Scrum Vorgehensweise

5.2. Künstliche Intelligenz im Projektmanagement

Künstliche Intelligenz (KI) kann große Datenmengen analysieren und Teams im Projektmanagement unterstützen. Sie hilft zum Beispiel bei der Priorisierung von Aufgaben, der Erkennung von Risiken oder der Analyse von Projektfortschritten. Tools wie Jira oder GitHub Copilot nutzen maschinelles Lernen, um Teams gezielt zu entlasten und Entscheidungen datenbasiert zu unterstützen [1] [4].

5.3. Die Kombination von KI und agilen Methoden

Agile Methoden und Künstliche Intelligenz verfolgen ähnliche Ziele: Teams sollen schneller, effizienter und flexibler arbeiten können. Die Kombination bietet die Chance, agile Prozesse durch datenbasierte Unterstützung weiter zu verbessern – zum Beispiel durch automatische Planungshilfen oder intelligente Risikoanalysen. Gleichzeitig stellt sie Unternehmen vor neue Herausforderungen, etwa bei der Integration in bestehende Teams und Prozesse [4] [1].

6. Methodik

Zur Beantwortung der Forschungsfrage wurde ein qualitatives Forschungsdesign gewählt. Ziel ist es, durch Experteninterviews praxisnahe Einblicke in den Einsatz von Künstlicher Intelligenz im agilen Projektmanagement zu gewinnen.
Auf Grundlage einer Literaturrecherche wurde ein halbstrukturierter Interviewleitfaden entwickelt. Die Befragung richtet sich an vier bis fünf Fachpersonen mit Erfahrung im agilen Arbeiten und im Umgang mit KI-gestützten Tools. Die Auswahl erfolgt mithilfe eines kurzen Screening-Bogens.
Die Interviews werden leitfadengestützt durchgeführt und dokumentiert. Die Fragen decken die Bereiche Praxiserfahrung, aktueller KI-Einsatz, Herausforderungen, Auswirkungen auf agile Arbeitsweisen sowie Zukunftseinschätzungen ab. Die Antworten werden thematisch zugeordnet, in einer Tabelle zusammengefasst und stichpunktartig ausgewertet.
Ziel der Analyse ist es, wiederkehrende Muster, abweichende Einschätzungen sowie konkrete Handlungsempfehlungen abzuleiten.

7. Ausblick

Die Verbindung von Künstlicher Intelligenz und agilen Methoden gilt als vielversprechender Ansatz, um Projektarbeit flexibler und datengestützter zu gestalten. Erste Entwicklungen zeigen, dass KI insbesondere in den Bereichen Planung, Entscheidungsunterstützung und Risikomanagement neue Möglichkeiten eröffnet.
Gleichzeitig fehlen in vielen Unternehmen noch klare Vorgehensmodelle, Erfahrungen im praktischen Einsatz und ein Verständnis dafür, wie sich KI sinnvoll in bestehende agile Strukturen integrieren lässt. Künftige Forschung sollte sich daher verstärkt der Frage widmen, unter welchen Bedingungen diese Technologien einen tatsächlichen Mehrwert liefern – sowohl für die Teams als auch für die Organisation.
Die im Rahmen dieser Arbeit geplanten Interviews sollen hierzu erste Erkenntnisse aus der Praxis liefern und konkrete Ansatzpunkte für eine gezielte Weiterentwicklung aufzeigen.

Literatur und Abbildungen

[1] Walter Brenner, Benjamin van Giffen, Jana Koehler, Tobias Fahse, and André Sagodi. *Bausteine eines Managements Künstlicher Intelligenz: Eine Standortbestimmung*. Springer Gabler, 2021.

[2] Eigene Darstellung.

[3] André Dechange. *Projektmanagement – Schnell erfasst*. Springer Gabler, 2024.

[4] Markus Glück. *Agile Innovation: Mit neuem Schwung zum Erfolg*. Springer Vieweg, 2025.

[5] David Schiefer. *Agile Skalierungsframeworks in der Theorie und Praxis: Einsatzgebiete und Grenzen im Vergleich*. Springer Gabler, 2022.

KI-gestützte Qualitätssicherung von UX-Konzepten am Beispiel von Figma und einer Webanwendung

Louis Asch **Dieter Morgenroth**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma entrecode GmbH, Stuttgart

Einleitung

Die Qualitätssicherung der User Experience und der damit verbundenen UX-Konzepte ist eine zentrale Anforderung in der modernen Softwareentwicklung. Während funktionale Anforderungen oft automatisiert überprüft werden können, bleibt die visuelle und strukturelle Konsistenz zwischen Design und Implementierung eine Herausforderung. Besonders in iterativen, agilen Prozessen entstehen durch kurzfristige Änderungen oder fehlende Dokumentation Abweichungen vom ursprünglichen Entwurf [1]. Designsysteme wie Figma ermöglichen eine strukturierte Planung von Benutzeroberflächen. Jedoch fehlen in der Praxis häufig automatisierte Verfahren zur systematischen Überprüfung, ob die final implementierte Webanwendung mit dem ursprünglichen Design übereinstimmt. In diesem Kontext bietet der Einsatz von Künstlicher Intelligenz vielversprechende Ansätze, um den visuellen Abgleich automatisiert und skalierbar umzusetzen [3].

Zielsetzung

Ziel dieser Arbeit ist es, ein Konzept inklusive Prototypen zu entwickeln, welche automatisiert Unterschiede und Diskrepanzen zwischen einem in Figma entwickelten UX-Konzept und einer dazugehörigen Webanwendung aufzuzeigen. Dabei sollen nicht nur Designabweichungen zwischen Konzept und Implementierung erfasst werden, sondern auch Diskrepanzen zwischen verschiedenen Versionen derselben Anwendung. Um dies zu erreichen, wird untersucht, wie KI-gestützte Methoden effizient eingesetzt werden können, um solche Abweichungen systematisch zu identifizieren. Ein weiterer Fokus liegt darauf, den Ansatz in den bestehenden Entwicklungsprozess zu implementieren. Zudem werden die Herausforderungen analysiert, die sich bei der automatisierten Erkennung von Änderungen in der Benutzeroberfläche ergeben.

Grounding DINO und SAM

Eine zentrale technische Grundlage dieses Ansatzes ist die Kombination der Deep-Learning-Modelle Grounding DINO und Segment Anything Model (SAM). Diese ermöglichen die automatisierte Erkennung und exakte Segmentierung semantisch relevanter UI-Komponenten aus Screenshots der Figma-Designs und der Webanwendung. Grounding DINO ist ein modernes Open-Set-Object-Detection-Modell, das visuelle Bildbereiche mit beliebig formulierten Textprompts verknüpfen kann. Anstatt auf feste Klassen trainiert zu sein, erkennt es mithilfe eines Sprach-Bild-Embeddings Elemente wie „CTA Button“ oder „Navigation Header“, selbst wenn diese in der Trainingsphase nicht enthalten waren [4]. Dies ist besonders vorteilhaft in flexiblen UI-Umgebungen, in denen Bezeichnungen variieren können. Abbildung 1 illustriert den grundlegenden Ablauf in Grounding DINO: Ein Bild und ein Textprompt werden durch getrennte Encoder verarbeitet, bevor ein Matching-Modul potenziell relevante Bildbereiche lokalisiert. Die erkannten Bounding Boxes dienen anschließend als Grundlage für die Segmentierung durch SAM.

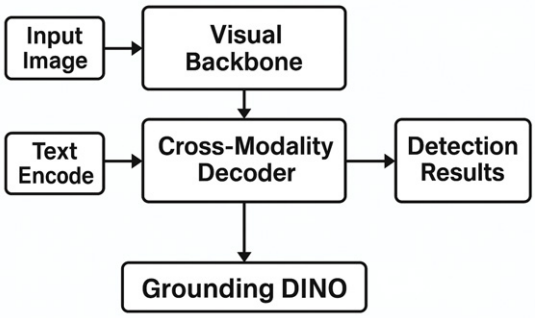


Abb. 1: Vereinfachte Darstellung des Grounding DINO Workflows [2]

Das SAM ergänzt Grounding DINO durch pixelgenaue Segmentierung. Das Modell wurde von Meta AI entwickelt, um Objekte im Bild unabhängig von Klassenzugehörigkeit zu maskieren. Mithilfe eines Prompt-Encoders, der Bounding Boxes verarbeitet, generiert SAM präzise Masken, die sich hervorragend zur Weiterverarbeitung in der Analyse und zum Bildvergleich eignen [3]. Die Kombination beider Modelle erlaubt es, UI-Komponenten wie Buttons, Bilder oder Karten auf semantischer Ebene zu erkennen und präzise auszuschneiden, unabhängig von ihrer Position oder Stilistik. Damit entsteht ein leistungsstarkes Werkzeug zur automatisierten und robusten UX-Vergleichsanalyse.

Ablauf

Zunächst werden über die Figma-API sämtliche Screenshots der UX-Konzepte extrahiert und anschließend durch einen Web-Scraper korrespondierende Screenshots der implementierten Webanwendung erzeugt. Diese Bilddaten durchlaufen anschließend eine Segmentierungspipeline, bestehend aus den Modellen Grounding DINO und SAM. Grounding DINO erkennt semantische UI-Komponenten wie Buttons, Bilder oder Navigationsleisten auf Basis vordefinierter textueller Prompts. Die erkannten Bounding Boxes werden an SAM übergeben, das daraus pixelgenaue Masken erzeugt.

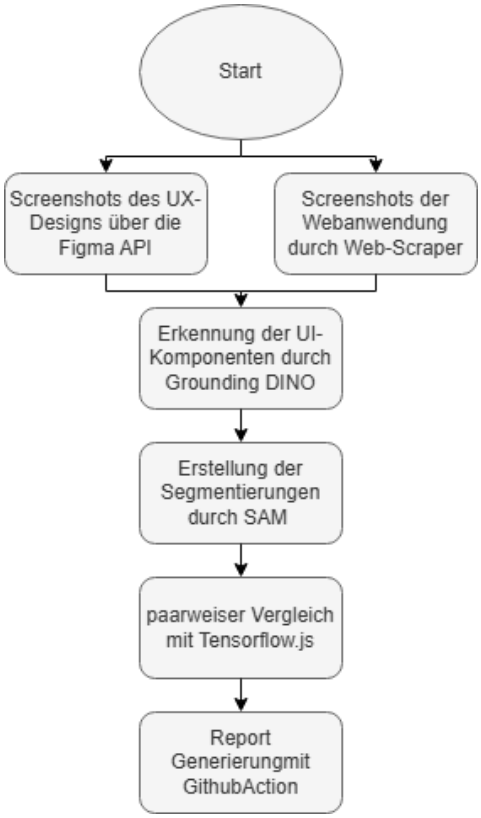


Abb. 2: Ablauf zum Erstellen des Reports [2]

Die segmentierten Bestandteile werden in standardisierter Form gespeichert und anschließend paarweise mit TensorFlow.js verglichen. Dabei werden Unterschiede auf Ebene der Pixel sowie über Bounding-Box-Ähnlichkeiten (IoU – Intersection over Union) und Label-Übereinstimmung bewertet. Die Ergebnisse werden anschließend als numerischer Differenzwert sowie als Visualisierung in einem automatisch generierten Report zusammengefasst. Die gesamte Ausführung wird über eine CI-Pipeline in GitHub Actions orchestriert.

Herausforderungen

Ein zentrales Problem stellt die Konsistenz der Darstellung dar. Webanwendungen variieren abhängig vom Ausgabegerät, Bildschirmauflösung oder Browser-Rendering. Um fehlerhafte Bewertungen zu vermeiden, muss diese Variabilität vor dem Vergleich durch Standardisierung oder Erkennung irrelevanter Unterschiede kompensiert werden. Zudem ist die semantische Zuordnung der richtigen Komponenten entscheidend. Die Übereinstimmung zwischen einem UI-Element aus dem Design und dessen Umsetzung in der Website basiert nicht nur auf Pixelähnlichkeit, sondern auch auf der semantischen Erkennung durch Grounding DINO. Uneindeutige Prompts, Mehrdeutigkeiten oder doppelt erkannte Elemente müssen daher herausgefiltert oder durch deduplizierende Nachverarbeitungsschritte aufgelöst werden.

Evaluation und Ausblick

Um die Wirksamkeit des vorgestellten Verfahrens zu beurteilen, wurden gezielt visuelle Änderungen an der bestehenden Webanwendung vorgenommen. Dabei zeigten sich Resultate bei größeren Änderungen an strukturellen UI-Komponenten wie Buttons, Navigationsleisten oder Hero-Elementen. Trotz der erreichten Ergebnisse bestehen noch Herausforderungen. Besonders die Sensitivität gegenüber kleinen Layout-Verschiebungen bei responsiven Designs kann zu falsch-positiven Meldungen führen.
Ein besonderer Vorteil der eingesetzten Methode besteht darin, dass sie unabhängig von festen Klassenzuweisungen funktioniert. Durch die textbasierte Steuerung mittels Prompts bleibt das System flexibel gegenüber variierenden Bezeichnungen und UI-Layouts, die sich im Kontext agiler Entwicklung häufig ändern. Dadurch stellt der Ansatz eine vielversprechende Lösung zur Qualitätssicherung in der Frontend-Entwicklung dar. Zwar kann er menschliche Review-Prozesse nicht vollständig ersetzen, jedoch gezielt ergänzen und beschleunigen. Insbesondere in CI/CD-Umgebungen bietet der vorgestellte Ansatz eine effektive Möglichkeit, die UX-Qualität langfristig sicherzustellen.

Literatur und Abbildungen

[1] Manuel Brhel, Hendrik Meth, Alexander Maedche, and Karl Werder. Exploring principles of user-centered agile software development: A literature re-view. *Information and Software Technology*, 61:163–164, 2015.

[2] Eigene Darstellung.

[3] Alexander Kirillov et al. Segment Anything. https://arxiv.org/abs/2304.02643, 04 2023.

[4] Shilong Liu et al. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. https://arxiv.org/abs/2303.05499, 07 2024.

Virtuelles Lernen mit KI-gestützten NPCs in Unreal Engine 5

Mehmet Kaan Asik **Dieter Morgenroth**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

Die Vorbereitung auf Prüfungen stellt für viele Studierende eine zentrale Herausforderung dar. Während klassische Lernmethoden wie das Lesen von Lehrbüchern oder Online-Skripten eine strukturierte Grundlage bieten, fehlt es diesen Formaten häufig an Interaktivität und persönlicher Ansprache. Dies führt nicht selten zu Motivationsverlust, oberflächlichem Lernen und einer ineffizienten Aufnahme des Lernstoffs [4].

Besonders Virtuelle Realität (VR) und Künstliche Intelligenz (KI) eröffnen innovative Wege, um Lernen nicht nur effizienter, sondern auch motivierender zu gestalten. VR ermöglicht es, Lerninhalte dreidimensional und immersiv zu erleben, wodurch das Gefühl von Präsenz gefördert und die kognitive Verarbeitung verbessert werden kann [5]. Gleichzeitig erlaubt KI, insbesondere durch leistungsfähige Sprachmodelle wie LLaMA, eine individuelle Anpassung von Lernprozessen sowie natürliche, dialogbasierte Interaktionen mit digitalen Tutoren [6].

Die Kombination beider Technologien kann neue Maßstäbe in der Hochschullehre setzen: Lerninhalte können nicht nur realistischer vermittelt, sondern auch unmittelbarer erfahren werden. VR erlaubt es etwa, durch Simulationen komplexe Prozesse (z. B. chemische Reaktionen oder physikalische Experimente) visuell verständlich zu machen. Die KI übernimmt dabei die Rolle eines personalisierten Tutors, der Rückfragen in natürlicher Sprache beantwortet, Quizfragen generiert oder Lernsituationen an das individuelle Verständnisniveau anpasst.

Didaktische Modelle wie das von Mayer (2001) [3] entwickelte Konzept des multimedialen Lernens belegen, dass die gleichzeitige Ansprache mehrerer Sinneskanäle – etwa visuell, auditiv und interaktiv – den Wissenserwerb nachhaltig verbessern kann. Diese Erkenntnisse legen nahe, dass die Kombination aus immersiver VR-Technologie und adaptiver KI das Potenzial besitzt, die Motivation sowie die Lernwirksamkeit deutlich zu steigern.

In dieser Arbeit wurde der Prototyp einer interaktiven VR-Anwendung entwickelt, die es Nutzerinnen und Nutzern ermöglicht, mit KI-gestützten NPCs (Non-Player Characters, also nicht vom Menschen gesteuerte Figuren) in einer virtuellen Lernumgebung zu interagieren. Ziel ist es, eine immersive und adaptive Lernplattform zu schaffen, in der digitale Tutoren individuell auf fachliche Inhalte zugreifen und auf Nutzereingaben in natürlicher Sprache reagieren können. Die technische Umsetzung kombiniert eine VR-Oberfläche mit einem intelligenten Backend, das PDF-Wissen vektorisiert, interpretiert und dialogfähig über ein Sprachmodell verfügbar macht.

Systemübersicht

Die folgende Abbildung veranschaulicht den Gesamtprozess des Systems – vom Nutzerinput über die Verarbeitung im Backend bis zur semantischen Beantwortung durch das KI-Modul.

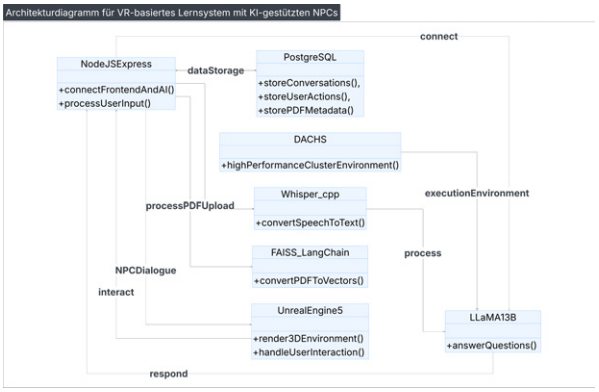


Abb. 1: Architekturdiagramm für VR-basiertes Lernsystem mit KI-gestützten NPCs [2]

Die folgende Abbildung zeigt die technische Ablaufstruktur des Systems – von der Sprach- oder Dateieingabe des Nutzers bis zur Beantwortung

durch das LLM. Sie veranschaulicht, wie verschiedene Module (z. B. Whisper.cpp, LangChain, FAISS, DACHS-Cluster) über das Backend miteinander verbunden sind.

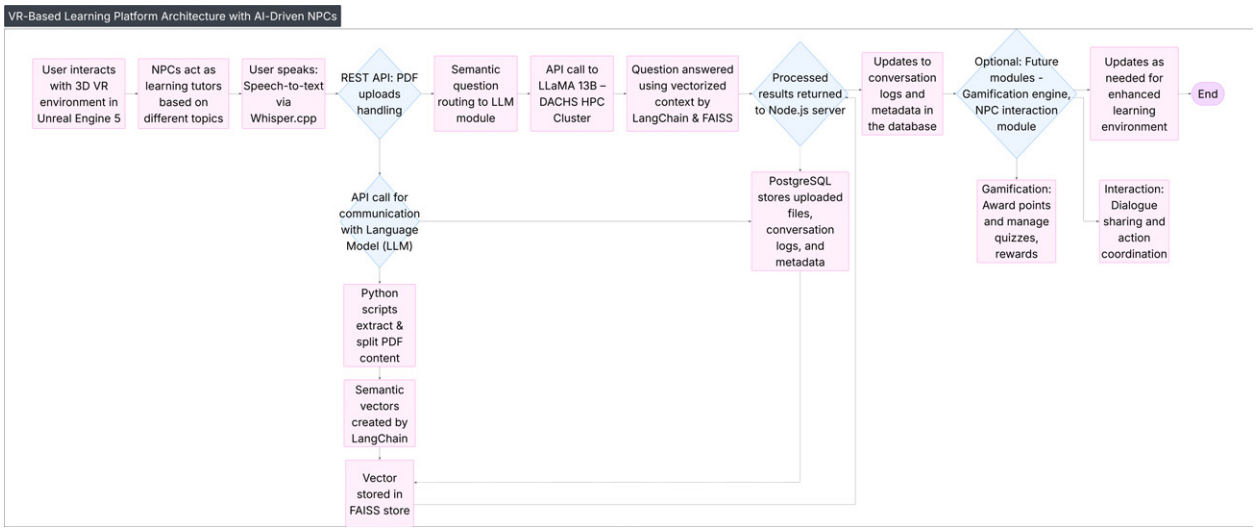


Abb. 2: Verarbeitungsfluss des VR-Lernsystems – inklusive Speech-to-Text, PDF-Upload, semantischer Vektorisierung, LLM-Anbindung und Datenbankmanagement.* [2]

Die Architektur basiert auf einem modularen Aufbau mit den folgenden Hauptkomponenten:

- Frontend: Die Benutzeroberfläche und VR-Interaktion erfolgen in Unreal Engine 5. Hier agieren die NPCs visuell und räumlich. Jede Figur kann einem bestimmten Fachgebiet zugewiesen werden und nutzt eine eigene Wissensbasis.
- Backend: Node.js mit Express dient als API-Server für Uploads, Anfragen an das LLM und das Management der Vektor-Daten. Zudem übernimmt das Backend die Steuerung von Logging, Nutzerverwaltung und ggf. Multiplayer-Kommunikation.
- KI-Modul: Das Sprachmodell LLaMA 13B wird nicht lokal, sondern auf dem Hochleistungscluster DACHS betrieben. Es verarbeitet Nutzeranfragen semantisch mit LangChain. Für jeden Kursbereich wird ein eigener Vektorindex erstellt.
- Datenbank: PostgreSQL speichert Konversationen, Uploads und Metadaten. Die Struktur umfasst Nutzersitzungen, Chatverläufe, Dateizuordnungen und Nutzungsstatistiken.

Funktionale Abläufe

Nutzer können PDF-Dateien hochladen, die anschließend mit LangChain in semantische Vektoren (mittels FAISS) überführt werden. Diese Inhalte stehen dann dem zuständigen NPC als Wissensbasis zur Verfügung. Fragen an den NPC werden semantisch interpretiert, relevante Textsegmente gesucht und anschließend durch das Sprachmodell beantwortet.
Die Kommunikation erfolgt entweder direkt (Text) oder über Spracheingabe (mittels Whisper.cpp). Zukünftig sollen auch Gamification-Elemente wie Quizze, Punktesysteme oder Belohnungen integriert werden, um die Motivation zusätzlich zu steigern. Diese Funktionen sollen den Lernprozess spielerisch begleiten und personalisierte Herausforderungen ermöglichen. Zudem ist geplant, die Interaktion mit digitalen Objekten zu ermöglichen – etwa durch das Platzieren von Molekülen oder das Lösen von Physikrätseln in der Umgebung.

Virtuelle Interaktion und Ausblick

In kommenden Entwicklungsschritten sollen die NPCs nicht nur dialogisch agieren, sondern auch physisch in der 3D-Welt handeln können. Durch den Einsatz von Blueprints und Interaktionslogik in Unreal Engine 5 wird es möglich sein, komplexe Abläufe in Echtzeit zu simulieren, wie z. B.:

- Schreiben an Whiteboards mit virtueller Handschrift
- Durchführung virtueller Laborexperimente
- Kollaborative Dialoge und Rollenspiele zwischen verschiedenen NPCs
- Bewegungen und Gestik mit synchronisierter Sprachausgabe (Text-to-Speech)

Eine eigene Online-Umfrage unter Studierenden [1] hat ergeben, dass 78 % der Befragten regelmäßig digitale Tools zum Lernen nutzen. 68 % zeigen großes Interesse an KI-gestützten Tutoren, und über 47 % würden gerne mit VR-Technologie lernen. Diese Ergebnisse belegen ein starkes Bedürfnis nach interaktiven, digitalen Lernformaten, die weit über klassische Skripte hinausgehen.

Fazit

Dieses Projekt demonstriert die Synergie zwischen immersiven VR-Technologien und fortschrittlicher Sprach-KI. Durch die Auslagerung des Sprachmodells auf den leistungsstarken DACHS-Cluster wird eine skalierbare und performante Lösung geschaffen. Die geplante Integration von Gamification sowie physischer Objektinteraktion verspricht ein nachhaltiges Lernerlebnis, das sowohl motiviert als auch didaktisch fundiert ist. Die Kombination moderner Technologie mit den explizit geäußerten Bedürfnissen der Zielgruppe legt den Grundstein für ein innovatives, zukunftsorientiertes Bildungssystem.

Literatur und Abbildungen

[1] M. Kaan Asik. Online-Befragung zur Akzeptanz von KI-basierten Lernsystemen. https://docs.google.com/forms/d/19HGIeOyQLD_DN4zmAoIWi8vgCSd1sdBRR-W989zQZZ8/edit, 03 2025.

[2] Eigene Darstellung.

[3] Richard Mayer. *Multimedia Learning*. Cambridge University Press, 2001.

[4] Jaziar Radianti, Tim A. Majchrzak, Jennifer Fromm, and Isabell Wohlgenannt. A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda. *Computers & Education*, page 147, 2020.

[5] Mel Slater and Sylvia Wilbur. A Framework for Immersive Virtual Environments (FIVE): Speculations on the Role of Presence in Virtual Environments. *Presence: Teleoperators and Virtual Environments*, 6:603–616, 1997.

[6] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marieanne Lachaux, Baptiste Roziere, et al. LLaMA 2: Open Foundation and Fine-Tuned Chat Models. https://arxiv.org/abs/2307.09288, 05 2023.

KI-gestütztes UX Design: Wie verbessern intelligente Systeme die Nutzungserfahrung auf Webseiten?

Sevde Aydin **Astrid Beck**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einführung

Die User Experience (UX) hat sich in den letzten Jahren zu einem entscheidenden Erfolgsfaktor für Webseiten entwickelt. Nutzer:innen erwarten nicht nur funktionierende Systeme, sondern personalisierte und sinnvolle Interaktionen. Die Integration von KI-Technologien wie Chatbots eröffnet neue Potenziale für personalisierte Nutzungserlebnisse und die dynamische Anpassung von Inhalten an das Nutzerverhalten [1].
UX lässt sich jedoch nicht allein durch technische Funktionalität erfassen. Vielmehr sind Nutzungserlebnisse das Resultat einer aktiven Auseinandersetzung mit einem Produkt in einem bestimmten Anwendungskontext. Sie entstehen aus dem Zusammenspiel von Wahrnehmung, Motivation, Emotion und Handlung und werden durch individuelle Bedürfnisse, Erwartungen und Ziele geprägt [3].

Ziel der Arbeit

Ziel dieser Arbeit ist es, den Einsatz von künstlicher Intelligenz im UX-Design von Webseiten zu untersuchen und zu analysieren, inwieweit intelligente Systeme zur Verbesserung der digitalen Nutzungserfahrung beitragen können. Im Fokus steht dabei die Frage, ob und wie Nutzer:innen den Einsatz von KI-basierten Funktionen wie Chatbots, Empfehlungssystemen oder KI-generierten Inhalten als Verbesserung oder auch als Erschwernis in der Interaktion mit Webseiten erleben. Neben den Potenzialen werden auch kritische Aspekte betrachtet, etwa fehlende Transparenz, kognitive Überforderung oder Vertrauensprobleme bei automatisierten Systemen. Die Betrachtung erfolgt interdisziplinär an der Schnittstelle von Informatik, Design und Nutzerpsychologie.
Zur Beantwortung der Fragestellung wird eine quantitative Nutzerumfrage durchgeführt, die Einblicke in Erwartungen, Erfahrungen und Einschätzungen im Umgang mit KI-gestützten Webseiten liefern soll. Erfasst werden dabei nicht nur Bewertungen bestehender Funktionen, sondern auch offene Wünsche, Kritikpunkte und Verbesserungsvorschläge aus Nutzersicht. Die Ergebnisse dienen als Grundlage für die Einordnung von Chancen und Grenzen intelligenter Systeme im UX-Kontext. Sie sollen praxisnahe Empfehlungen ermöglichen, sowohl zur Optimierung bestehender als auch zur Entwicklung neuer, nutzerzentrierter Webanwendungen.

KI-Technologien im UX-Kontext

Künstliche Intelligenz ist heute ein zentraler Bestandteil digitaler Anwendungen. Auch im Bereich der UX spielt sie eine zunehmend wichtige Rolle. Besonders relevant ist sie dort, wo Prozesse automatisiert, Inhalte personalisiert oder Interaktionen vereinfacht werden sollen. Typische Technologien sind Chatbots, Empfehlungssysteme und generative Modelle zur Erzeugung von Texten, Bildern oder Elementen der Benutzeroberfläche (User Interface, UI) [4]. Die in diesem Zusammenhang relevanten Technologien sind in Abbildung 1 dargestellt.

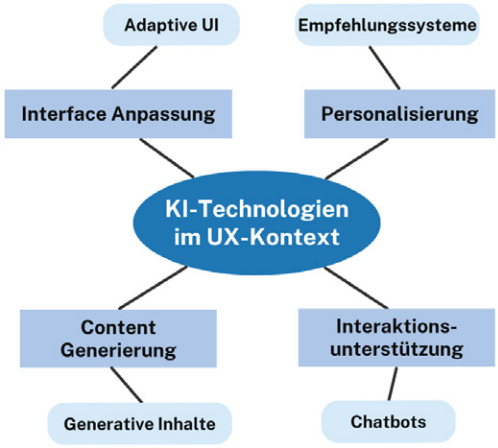


Abb. 1: KI-Technologien im UX-Kontext [2]

Ein weiterer Aspekt, der im Kontext KI-gestützter UX-Designs zunehmend an Bedeutung gewinnt, ist die

Qualität der zugrunde liegenden Daten. Personalisierte Erlebnisse und präzise Empfehlungen sind nur dann möglich, wenn die Systeme auf umfangreiche, aktuelle und qualitativ hochwertige Daten zurückgreifen können [4]. Verzerrte oder unvollständige Datensätze können hingegen zu fehlerhaften Entscheidungen führen und das Vertrauen der Nutzer:innen in KI-gestützte Funktionen beeinträchtigen. Diese Systeme nutzen meist Verfahren des maschinellen Lernens. Häufig kommen dabei Deep-Learning-Architekturen wie Foundation Models zum Einsatz. Bekannte Beispiele sind Large Language Models (LLMs) wie GPT-4. Sie erzeugen kontextabhängige Inhalte, zum Beispiel für Kundenchats oder automatisierte E-Mails [4].
Moderne Chatbots arbeiten nicht mehr mit festen Entscheidungsbäumen. Sie verwenden Natural Language Processing (NLP) und greifen auf sogenannte Dialogkorpora zurück [5]. Diese enthalten gesammelte Gesprächsdaten zwischen Mensch und Maschine.
Aus Sicht der UX ist nicht die technische Komplexität entscheidend, sondern die Wirkung auf das Nutzungserleben. Nutzer:innen beurteilen KI-Elemente oft nach Verständlichkeit, Personalisierung und Kontrollgefühl. Besonders bei generativen Inhalten oder automatisierten Empfehlungen sind Transparenz und Autonomie wichtige Faktoren. Sie beeinflussen, ob ein System als hilfreich oder störend empfunden wird [3].

Fazit und Ausblick

Die Integration von KI-Technologien bietet große Potenziale, um die Nutzungserfahrung auf Webseiten individueller und effizienter zu gestalten. Gleichzeitig stellen intelligente Systeme Designer:innen vor neue Herausforderungen. Nutzer:innen bewerten KI-Funktionen nicht nur anhand ihrer technischen Leistungsfähigkeit, sondern insbesondere im Hinblick auf Verständlichkeit, Transparenz und die Wahrung von Kontrolle. Die zentralen Einflussfaktoren, die dabei eine entscheidende Rolle spielen, sind in Abbildung 2 zusammengefasst.

Abb. 2: Einflussfaktoren auf die Nutzerwahrnehmung von KI-Anwendungen [2]

Die noch laufende Nutzerumfrage wird weitere Erkenntnisse darüber liefern, wie diese Aspekte in der Praxis wahrgenommen werden und welche Erwartungen an KI-gestützte Funktionen bestehen. Auf Basis der Ergebnisse sollen konkrete Empfehlungen für die nutzerzentrierte Gestaltung von Webseiten mit KI-Anwendungen entwickelt werden. Die Ergebnisse sollen nicht nur aktuelle Handlungsempfehlungen liefern, sondern auch einen Beitrag zur Gestaltung zukunftsfähiger, vertrauensvoller Mensch-KI-Interaktionen leisten.
Besonders für Unternehmen, die ihre digitalen Angebote strategisch weiterentwickeln möchten, bietet der gezielte Einsatz von KI-gestützten UX-Methoden einen wichtigen Wettbewerbsvorteil. Gleichzeitig bleibt die Herausforderung bestehen, technologische Innovationen mit den Erwartungen und Bedürfnissen der Nutzer:innen in Einklang zu bringen, um langfristig Akzeptanz und Vertrauen zu sichern. Neben diesen gestalterischen und nutzerzentrierten Anforderungen stellen auch technische Faktoren wie die Skalierbarkeit und Leistungsfähigkeit von KI-Systemen im Echtzeiteinsatz eine Herausforderung dar. Gerade bei der Integration komplexer Modelle müssen sowohl die technische Infrastruktur als auch wirtschaftliche Aspekte berücksichtigt werden.

Literatur und Abbildungen

[1] Guendalina Caldarini, Sardar Jaf, and Kenneth McGarry. A Literature Survey of Recent Advances in Chatbots. *Information*, 13, 2022.

[2] Eigene Darstellung.

[3] Sarah Diefenbach and Marc Hassenzahl. *Psychologie in der nutzerzentrierten Produktgestaltung: Mensch–Technik–Interaktion erleben*. Springer Vieweg, 1 edition, 2017.

[4] Bernhard Wecke. *Wachstum durch den Einsatz von Generativer KI: Funktionsweise und Anwendungsgebiete im Marketing*. Springer Essentials, 1 edition, 2024.

[5] Gregor Wilke and Oliver Bendel. KI-gestütztes Recruiting – technische Grundlagen, wirtschaftliche Chancen und ethische Herausforderungen. *HMD – Praxis der Wirtschaftsinformatik*, 59, 2022.

Konzeption eines KI-basierten Multi-Agenten-Systems für Geschäftsprozessverbesserungen: Ein Design-Science-Research-Ansatz

Dominik Bajrami **Manfred Schoch**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Fraunhofer-Institut für Angewandte Informationstechnik FIT, Augsburg

Einleitung

Unternehmen müssen ihre Marktposition und Geschäftsabläufe kontinuierlich hinterfragen, da wirtschaftlicher Wandel und technologische Entwicklungen permanenten Anpassungsdruck erzeugen. Um unter diesen Bedingungen erfolgreich agieren zu können, müssen Unternehmen in der Lage sein, sich kontinuierlich weiterzuentwickeln. Die Geschäftsprozesse und deren IT-Unterstützung bilden dabei eine zentrale Grundlage der Wertschöpfung [1]. Eine zentrale Rolle spielt hierbei die Optimierung von Geschäftsprozessen. Business Process Improvement (BPI) hilft dabei bestehende Prozesse effizienter, qualitativ hochwertiger und anpassungsfähiger zu gestalten.

Problemstellung und Zielsetzung

Während die systematische Analyse der bestehenden Prozesse bereits in großem Umfang durch Process Mining automatisiert werden kann, erfolgt die Optimierung und Neugestaltung momentan überwiegend manuell, was sehr viele Ressourcen in personeller und zeitlicher Hinsicht in Anspruch nimmt. Diese Verfahren sind nicht nur kostenintensiv, sondern auch stark von individuellen Fähigkeiten und Erfahrungen von Mitarbeitenden abhängig. Generative künstliche Intelligenz kann helfen, diese Limitationen zu überwinden. Ziel dieser Arbeit ist es daher, ein Konzept für ein System zu entwickeln, das in der Lage ist, diese Lücke zu schließen. Es soll ein auf künstlicher Intelligenz (KI) basiertes Multi-Agenten-System konzipiert werden, dass durch den Einsatz generativer KI eigenständig kreative Prozessverbesserungsvorschläge generieren kann. Dabei sollen spezialisierte Agenten unterschiedliche Rollen einnehmen, wie z. B. kreative Ideengenerierung, Prozessanalyse, Prozessbewertung, oder Feedbackverarbeitung.

Theoretische Grundlagen

Geschäftsprozesse stellen die Grundbausteine jeder wertschöpfenden Tätigkeit eines Unternehmens dar. Sie beschreiben eine Abfolge von Tätigkeiten, in deren Verlauf eine oder mehrere Arten von Eingaben verarbeitet werden, um ein Ergebnis zu erzeugen, das einen Mehrwert generiert [3]. Geschäftsprozesse werden zunehmend durch digitale Technologien unterstützt und sind ein zentraler Hebel für unternehmerischen Erfolg. Business Process Improvement ist ein methodischer Ansatz, der Unternehmen gezielt dabei unterstützt, ihre bestehenden Geschäftsprozesse zu verbessern. Mit diesem Ansatz können Prozesse klarer strukturiert und effizienter gestaltet werden [4]. Beim Business Process Reengineering, steht die komplette Neubewertung bestehender Prozesse im Mittelpunkt. Dabei wird nicht von dem ausgegangen, was derzeit vorhanden ist, sondern es wird überlegt, wie der Prozess aussehen würde, wenn er auf der Basis des heutigen Wissens und der heutigen technologischen Möglichkeiten völlig neu entwickelt würde [5].

Methodik: Design Science Research (DSR)

Zur Beantwortung der Forschungsfrage wird ein Design-Science-Research-Ansatz verfolgt. Dieser Forschungsansatz eignet sich besonders zur Gewinnung gestaltungsorientierter Erkenntnisse und zur Entwicklung neuer IT-Artefakte - in diesem Fall einer Systemarchitektur. Die Arbeit gliedert sich in folgende methodische Schritte [6] , wie in Abbildung 1 zu sehen ist:

1. Problemidentifikation und Zieldefinition
2. Ableitung von Designanforderungen
3. Modellierung einer Referenzarchitektur

4. Konzeptionelle Evaluation des Modells anhand theoretischer Kriterien

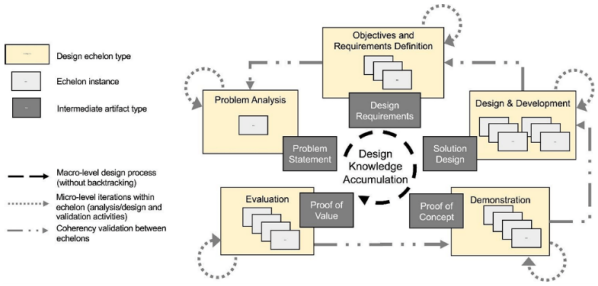


Abb. 1: eDSR-Metamodell [6]

Die Anforderungen wurden aus einer Literaturrecherche abgeleitet und in fünf Design Objectives (DO1–DO5) gegliedert. Diese umfassen unter anderem die Fähigkeit zur Generierung kreativer Ideen, die semantische Integration externer Wissensquellen, die Perspektivenvielfalt durch spezialisierte Agentenrollen, sowie Transparenz- und Bewertungsfähigkeit.

Aktueller Stand der Architektur

Das konzipierte Multi-Agenten-System, wie in Abbildung 2 zu sehen ist, sieht eine modulare Struktur vor.

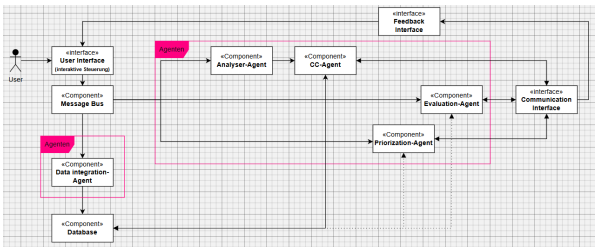


Abb. 2: Erste konzeptionelle Darstellung der Systemarchitektur [2]

Verschiedene Agenten übernehmen dabei spezialisierte Aufgaben, etwa:

- **Computational Creativity-Agent:** Verantwortlich für die Generierung von Prozessverbesserungsvorschlägen.
- **Analyser-Agent:** Identifiziert und analysiert die Verbesserungspotenziale auf Basis der vorhandenen Ist-Prozesse und Kontextinformationen.
- **Evaluation-Agent:** Bewertet die zuvor erzeugten Vorschläge anhand definierter Leistungsindikatoren und strategischer Zielkriterien.
- **Priorization-Agent:** Priorisiert die bewerteten Vorschläge gemäß ihrer erwarteten Wirkung und Relevanz für die Unternehmensziele.
- **Data-Integration-Agent:** Aggregiert und harmonisiert strukturierte sowie unstrukturierte Wissensquellen, um eine kontextbasierte Entscheidungsgrundlage für die anderen Agenten bereitzustellen.

Über ein User Interface erfolgt die Eingabe durch den Nutzenden, welche anschließend über den Message Bus zur weiteren Verarbeitung innerhalb des Systems überführt wird. Das Feedback Interface dient der strukturierten Erfassung von Nutzerfeedback, das in den iterativen Verbesserungsprozess der generierten Vorschläge einfließt. Die Kommunikation zwischen den einzelnen Agenten wird über ein zentrales Communication Interface realisiert, das sowohl den Informationsaustausch ermöglicht als auch die Koordination der Interaktionen übernimmt. Diese Architektur ist bewusst offengehalten und erlaubt eine schrittweise Erweiterung um weitere Agenten oder Technologien.

Ausblick

Im weiteren Verlauf der Arbeit wird die Referenzarchitektur detailliert ausgearbeitet und ggf. einer theoretischen Evaluation durch Experteninterviews unterzogen. Ziel ist es, zu beschreiben, wie ein KI-basiertes Multi-Agenten-System gestaltet werden kann, um eigenständig kreative Prozessverbesserungsvorschläge zu erzeugen. In zukünftigen Arbeiten könnte ein prototypischer Implementierungsversuch folgen, um die praktische Umsetzbarkeit zu erproben. Langfristig ermöglicht die Kombination generativer KI und Multi-Agenten-Systeme die Automatisierung kreativer Managementaufgaben.

Literatur und Abbildungen

[1] Jörg Becker, Christoph Mathas, and Axel Winkelmann. *Geschäftsprozessmanagement*. Springer Berlin, Heidelberg, 1 edition, 2009.

[2] Eigene Darstellung.

[3] Michael Hammer and James Champy. *Reengineering the Corporation: A Manifesto for Business Revolution*. HarperCollins, 2002.

[4] H. James Harrington. *Business Process Improvement: The breakthrough strategy for Total Quality, Productivity and Competitiveness*. McGraw-Hill, Inc., 1991.

[5] Rupert Hierzer. *Prozessoptimierung 4.0: Den digitalen Wandel als Chance nutzen*. Haufe-Lexware GmbH & Co. KG, 2 edition, 2020.

[6] Tuure Tuunanen, Jan vom Brocke, and Robert Winter. Dealing with Complexity in Design Science Research: A Methodology Using Design Echelons. *MIS Quarterly*, 48:427–458, 2024.

Optimizing Place Recognition in ORB-SLAM3 for Robotic Lawn Mowers

Maxim Becht **Thao Dang**

Department of Computer Science and Engineering, Esslingen University

Work carried out at Department of Computer Science and Engineering, Esslingen

Introduction

Navigating an autonomous lawn mower with great precision in outdoor environments using visual Simultaneous Localization and Mapping (SLAM) is a challenging endeavor. Many factors, such as sensor noise, lighting, and environmental conditions, can negatively impact drift build-up, as is the case in dynamic outdoor environments. To counteract drift build-up, visual SLAM often relies on loop closure, utilizing place recognition techniques [2]. Lawn mowers typically operate more efficiently when following parallel paths. During the autonomous mower learning phase, the outer limits of the environment are especially identified. Therefore, given the traveled trajectories observed during the initial and operating phases, place recognition in ORB-SLAM3 [2] is addressed in the context of changes in the horizontal perspective.

The main contribution of this paper is a novel optimization process to improve the place recognition of ORB-SLAM3. For the optimization process, a perspective-based dataset was recorded in a challenging outdoor environment, to evaluate the place recognition abilities of ORB-SLAM3. The optimization process involves the open source hyperparameter optimization framework Optuna [1] for parameter optimizations and image post-processing, such as brightness, contrast, and sharpness applied to pre-recorded simulation data. In addition, a custom visual words vocabulary is trained on sampled datasets from [4]. This work demonstrates that significant improvements in the place recognition capabilities of ORB-SLAM3 can be achieved through targeted parameter optimization, emphasizing the strong interdependence of the front and back end. Post-processing of input images notably enhances tracking in low-feature, dynamic environments, though it introduces a dependency on consistent processing between mapping and tracking stages.

The study also highlights an Inertial Measurement Unit (IMU) initialization delay in ORB-SLAM3, which, while ensuring map maturity, can hinder rapid place recognition early in mapping.

Final evaluations show that, with optimization, a horizontal perspective change of up to 30 degrees can be reliably tolerated, significantly improving the robustness of robotic lawn mowers operating under varying paths.

Custom vocabulary training with a small dataset (1,000 images) proved ineffective compared to the default generalized ORB-SLAM3 vocabulary, underlining the need for large and diverse training data.

For practical deployment, using predefined parameter sets based on environment and weather conditions appears promising for real-time systems, although full adaptive tuning remains challenging. However, further research is required on this topic.

Place Recognition in ORB-SLAM3

This section delves into the place recognition mechanism within the ORB-SLAM3 framework. Relocalization is treated as a multiphase process that is dissected into three main stages: Transformation Finding, Optimization, and Verification. These stages are designed to determine whether a camera revisits a known location, using visual and inertial cues within the system's map structure. The overall workflow is based on both the DBOW2 place recognition algorithm and code-level behavior of ORB-SLAM3, as well as remarks from [2].

Stage 1: Find 3D Transformation

The relocalization process initiates once the loop closing thread accumulates at least one keyframe in its queue, which is checked every 5 milliseconds. When the system has initialized its IMU (i.e., after at least 12 keyframes and completion of the initial inertial bundle adjustment), a DBOW2 query is triggered on the active keyframe. This query returns the top three candidates based on appearance similarity and covisibility constraints. If a match is found in another map within the atlas, a map merge is attempted; otherwise, loop closing proceeds.

Each of these top matches leads to a search for 10 covisible keyframes to form local windows. Any local window too close to the active keyframe—defined by direct covisibility—is discarded. Valid local windows must share at least 10 map points with the active keyframe to proceed, as controlled by the nBow-Matches parameter.
Subsequently, a 3D transformation is computed using RANSAC. The algorithm selects sets of three 3D-3D correspondences and applies the Horn method to find the optimal rotation and translation. The best hypothesis is chosen based on reprojection error evaluations and vote counts. Successfully estimating this transformation marks the completion of the first stage.

Stage 2: Optimization

With the initial transformation in place, Stage 2 begins by refining this estimate through bidirectional matching. This includes not only projecting local map points into the active keyframe, but also the reverse. The search utilizes reprojection windows around keypoints, with matches validated using Hamming distances between feature descriptors.
A minimum number of inliers must be established to continue. Failing this, the process reverts to Stage 1. Upon finding adequate correspondences, optimization is performed using bidirectional reprojection error and stabilized by applying a Huber loss function to mitigate outlier effects. This optimization process is repeated once more with a tighter search window to enhance accuracy and confidence.

Stage 3: Verification

The final stage assesses the validity of the place recognition hypothesis. Unlike earlier stages that depend on global DBOW2 queries, verification selects test candidate keyframes from the active map based on covisibility with the current active keyframe. This localized strategy, a novel feature in ORB-SLAM3, enhances robustness—especially in visual-inertial and multi-map contexts—by emphasizing spatial continuity over visual similarity alone.
The verification proceeds until three keyframes confirm the recognition through sufficient matches or two consecutive candidates fail. Only after this verification step passes can the system proceed with merging map structures, ensuring both spatial consistency and multi-threaded safety.

Recording a perspective-based dataset

Capturing the perspective-based data set requires the use of a moving test platform that resembles the application device in question. The movable platform in this case is the Stihl iMOW® 4 mower, on which a T265 intelisense stereo camera with integrated IMU is mounted. For the recording of the ground truth data, the Leica total station TS16 is utilized. This total station is capable of tracking a mini-prism (Leica GRZ101) over an Automatic Target Recognition (ATR) range of 350m and a full 360° reflection angle coverage. A Network Time Protocol (NTP) server is hosted on the recorder device to synchronize the ground truth with the camera and the IMU data. The ground truth data is send from the stationary part of the recording setup to the recording device located on the testing platform over User Datagram Protocol (UDP).

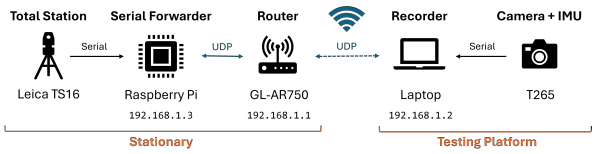


Fig. 1: Recording setup showing the primary components, categorized into stationary and moving equipment. [3]

With the recording setup as described and shown in Figure 1, it is possible to capture the ground truth trajectories of each angular run as shown in the top-down view in Figure 2. For the lawn mower to move in these preplaned routes, the "back home" functionality of the iMOW® 4 mower is utilized, which sends the mower back to the loading station by following the guarding/guiding wire.
The junction, marked in orange, where all angles meet, is crucial in the optimization process, as it marks the point until place recognition may be executed under the influence of perspective disparity. This event is referred to here as the "Wire Junction Reached" event.

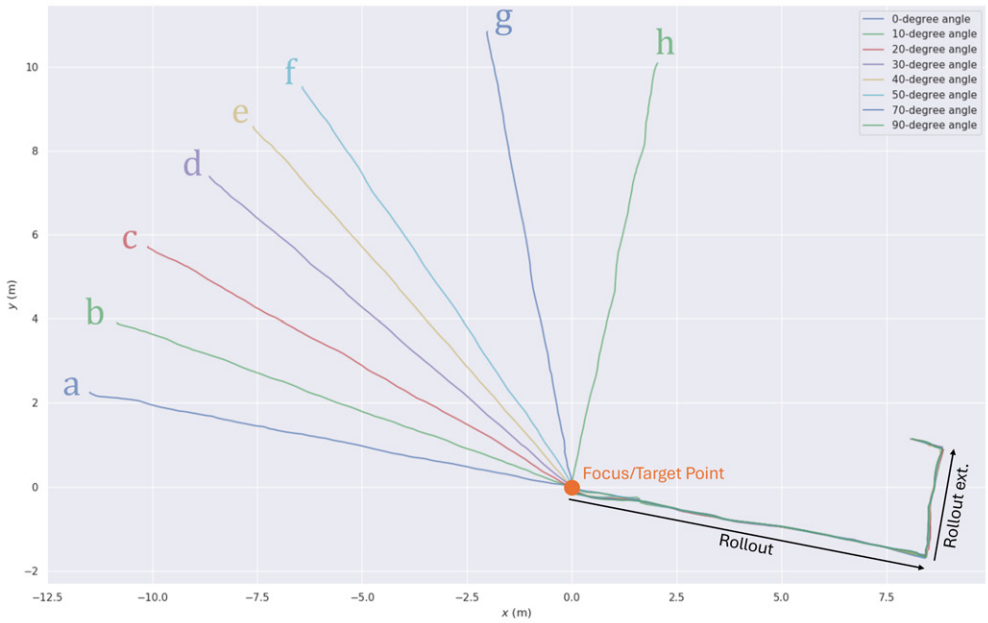


Fig. 2: Top down view of the combined ground truth trajectories [3]

Parameter Optimization

Optuna is a hyperparameter optimization framework that automates the search for optimal configurations using a defined-by-run approach and advanced sampling strategies, such as Tree-structured Parzen Estimators (TPE). In this study, Optuna is applied to optimize key front-end and back-end parameters of ORB-SLAM3 in a simulated environment. The optimization objective centers on minimizing RMSE for Relative and Absolute Pose Error (RPE, APE) while maximizing a custom relocalization metric, the proposed Weighted Relocalization Response Time (WRRT). WRRT quantifies the timing of map merging relative to a predefined event (wire junction reached), favoring early merges indicative of successful relocalization under perspective disparity. Two WRRT formulations are evaluated: a discontinuous Heaviside-based function and a smooth exponential variant, as defined in Equation 1 and 2 respectively. The latter mitigates artificial discontinuities in the parameter space, enabling more stable and representative optimization.

$$f(t) = t\,(0.5 + 0.5H(t - t_{\text{junction}})) \tag{1}$$

$$f(t) = a \cdot t + b \cdot e^{c \cdot (t - t_{\text{junction}})} \tag{2}$$

Post-Processing

During testing, multiple image adjustments were applied to the T265 images of the rosbag prior to the evaluation process. The direction of processing effects applied is the following:

1. Brightness
2. Contrast
3. Sharpness
4. Gaussian Blur
5. Canny Edge Detector

The brightness modification is implemented by uniformly increasing all pixel values (across all RGB channels) by an integer value. Clipping values outside the 0-255 range ensures that the image remains in a valid format.
The contrast adjustment is achieved by multiplying the image pixel values by a factor alpha, where values greater than 1 expand the range of pixel values, and values less than 1 compress it.
Sharpening is applied by convolving the image with a kernel, as defined by Equation 3, that emphasizes the center pixel, with the kernel values adjusted by the sharpening level L, enhancing edges and details.

$$\text{Sharpening Kernel} = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 5+L & -1 \\ 0 & -1 & 0 \end{bmatrix} \tag{3}$$

The Gaussian blur filter is applied before the Canny Edge Detector, to help minimize the impact of noise. By smoothing the image, the Gaussian blur reduces small variations and high-frequency noise that might cause false edges. Therefore, the accuracy of the Canny Edge Detector in finding edges is increased, reducing false positives. The Canny edge detector is a widely used image processing algorithm that detects edges

by optimizing three criteria: detection, localization, and minimal response. It applies Gaussian smoothing, gradient calculation, non-maximum suppression, and edge tracking by hysteresis. The idea behind using a canny edge detector before trying to extract ORB is to drastically increase the strong edges in the texture of the grass and to separate them aggressively from the weaker ones, completely filtering them.

Vocabulary Training

To enhance place recognition performance in ORB-SLAM3, a custom Bag of Visual Words (BoW) vocabulary was trained using environment-specific imagery from garden settings. The vocabulary, built with the DBOW2 library, encodes ORB binary descriptors into discrete visual words. A tailored dataset was compiled from rosbags captured using the same camera employed in SLAM simulations, ensuring consistency in image characteristics. To overcome the memory limitations of DBOW2, which requires all images to be loaded simultaneously, an automated ORB-based algorithm was used to sparse image sequences. This algorithm compares ORB descriptors between successive images, iteratively removing redundant frames through staged similarity thresholding and a final exhaustive comparison. The resulting dataset preserves high feature uniqueness while minimizing image count, covering six garden locations under diverse seasonal, lighting, and weather conditions. This curated dataset supports the offline training of a robust vocabulary optimized for the target deployment environment.

Results

Optimizing both the frontend and backend components of the ORB-SLAM3 system leads to substantial improvements in accuracy, robustness, and overall system performance, particularly in challenging dynamic environments.

Frontend Enhancements

Enhancing the frontend of the ORB-SLAM3 system through image post-processing and feature detector parameter tuning significantly improves tracking and relocalization in environments with low texture and sparse features. The PBCS method, as shown in Figure 3, enhances stereo matching by making subtle elements, such as grass, more detectable, which is especially beneficial when initializing the system in open grass fields with sparse vegetation and distant landmarks. Although Canny edge detection reduces stereo compatibility, PBCS consistently yields better map point generation and more stable tracking.

A key insight is the importance of alignment between map generation and inference. Maps processed with PBCS are not always compatible with rosbags that were not processed with PBCS, and vice versa. Moreover, frontend improvements alone are insufficient to guarantee successful relocalization; challenges such as IMU initialization and map compatibility remain significant. The results of frontend optimization are shown in Table 1 of 4.

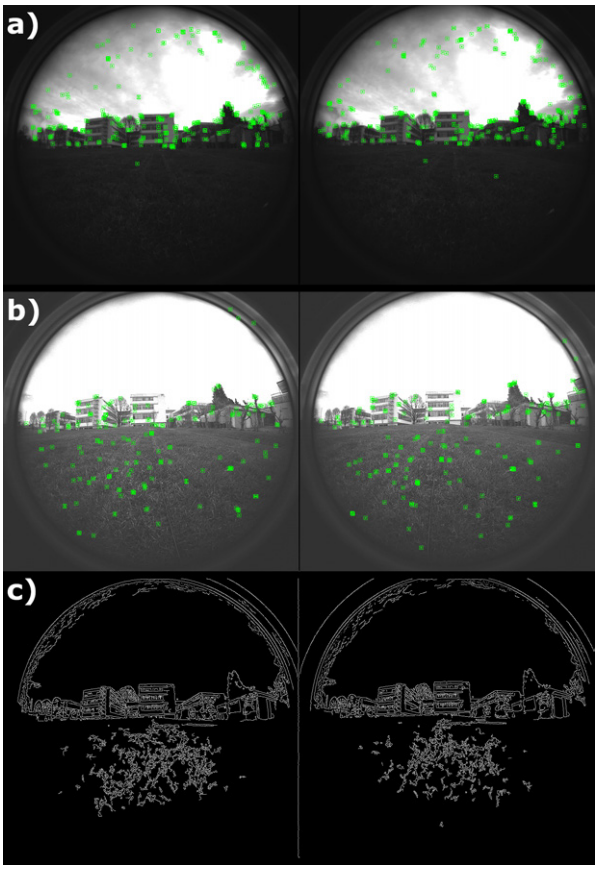


Fig. 3: Post-processing shown side by side with tracked stereo matches marked in green: (a) Default, (b) PBCS, and (c) Canny Edge Detector. All frontend samples were captured from the same zero-degree angle run. [3]

Backend Enhancements

Tuning the backend parameters of ORB-SLAM3 significantly enhances accuracy, robustness, and the success rate of map merging, especially when used in conjunction with an optimized frontend. While some parameter configurations may lead to fewer valid runs, those that succeed often achieve notably higher accuracy, as can be seen in Table 2 of 4.

During the early stages of map construction, place recognition performance may be improved by removing the VIBA condition. In ORB-SLAM3, the VIBA (Visual-Inertial Bundle Adjustment) condition requires the system to complete initialization, aligning visual and inertial data, before place recognition and loop

closure are enabled. While this ensures consistent and accurate mapping, it can delay place recognition when the map is still being built.
However, observed outlier behaviors and sensitivity to trajectory angles highlight trade-offs that make backend optimization complex, requiring adaptive parameter tuning for real-world deployment. Since implementing fully adaptive tuning across diverse environments is challenging, a practical compromise would be to create a look-up table of predefined parameter sets. These could be dynamically selected based on the environmental conditions detected by the system.

Table 1: Default relocalization parameters with optimized front-end parameters.
The initial map loaded in all cases is build by one run of the 0° (M) rosbag.

Rosbag		0° (M)	0°	10°	20°	30°	40°	50°	70°	90°
APE	Mean	2.046	3.187	4.544	1.437	0.155	2.545	2.960	3.255	2.417
	SD	1.583	1.918	1.834	2.523	0.026	2.349	1.396	0.989	1.446
RPE	Mean	0.518	0.165	0.139	0.040	0.022	0.112	0.110	0.110	0.102
	SD	0.427	0.064	0.126	0.044	0.006	0.162	0.101	0.083	0.021
Valid	%	70%	70%	80%	60%	30%	70%	80%	60%	40%
Merge	Mean	35.980	X	39.000	X	48.502	X	X	X	X
	SD	0.000	X	0.000	X	4.259	X	X	X	X
Merge relative	Mean	4.920	X	1.600	X	-7.0015	X	X	X	X
	SD	0.000	X	0.000	X	4.259	X	X	X	X
Valid	%	10%	0%	10%	0%	20%	0%	0%	0%	0%

Table 2: Optimized relocalization with manually optimized front-end parameters, adjusted based on multiple optimization results and observations during testing.
The initial map loaded in all cases is build by one run of the 0° (M) rosbag.

Rosbag		0° (M)	0°	10°	20°	30°	40°	50°	70°	90°
APE	Mean	2.330	2.277	2.446	1.407	0.229	1.651	2.296	2.547	2.250
	SD	1.190	2.215	2.143	1.433	0.260	1.994	1.747	0.984	1.622
RPE	Mean	0.279	0.192	0.178	0.240	0.050	0.468	0.413	0.234	0.225
	SD	0.296	0.139	0.089	0.200	0.055	0.623	0.499	0.044	0.174
Valid	%	80%	70%	90%	100%	90%	80%	80%	80%	70%
Merge	Mean	33.953	23.304	40.361	43.423	40.878	41.939	51.020	51.681	52.722
	SD	19.710	0.777	16.889	10.789	8.330	3.904	4.726	10.958	3.315
Merge relative	Mean	6.947	13.296	0.239	-2.323	0.6217	-3.539	-9.720	-11.181	-9.122
	SD	19.710	0.777	16.889	10.789	8.330	3.904	4.726	10.958	3.315
Valid	%	100%	30%	60%	40%	100%	100%	40%	90%	60%

Fig. 4: Optimization results of optimized frontend (Table 1) and front/backend (Table 2). [3]

Conclusion

This work explored improvements to place recognition in autonomous robotic lawn mowers using ORB-SLAM3, with a focus on addressing horizontal perspective changes common in outdoor navigation. A custom dataset was recorded to evaluate the system under perspective variation, and image post-processing was applied to enhance feature extraction in low-texture environments.
Parameter optimization via Optuna, targeting WRRT and RPE, significantly improved both recognition speed and accuracy. However, the benefits were found to be context-dependent, requiring consistent preprocessing between mapping and tracking. While a custom vocabulary trained on a small dataset underperformed, the default ORB-SLAM3 vocabulary proved robust for stereo fisheye cameras.
Results showed reliable place recognition for perspective changes up to ~30°, given proper tuning. Limitations include IMU initialization delays affecting early recognition and the challenge of implementing real-time adaptive tuning. Predefined parameter sets may offer a practical alternative, with future work needed to support dynamic adaptation.

References and figures

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A Next-generation Hyperparameter Optimization Framework. In *KDD '19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Association for Computing Machinery, New York, NY, United States, 2019.

[2] Carlos Campos, Richard Elvira, Juan J Gómez Rodríguez, José M M Montiel, and Juan D Tardós. ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual–Inertial, and Multimap SLAM. In *IEEE Transactions on Robotics*, volume 37, pages 1874–1890. IEEE, 2021.

[3] Own representation.

[4] Fabian Schmidt, Constantin Blessing, Markus Enzweiler, and Abhinav Valada. ROVER: A Multi-Season Dataset for Visual SLAM. In *arXiv preprint arXiv:2412.02506*. arXiv.org, 2024.

Konzeptentwicklung zur Modellierung von Zubehör- und Ersatzteilverkäufen basierend auf Produktabsätzen

Michael Becker **Markus Enzweiler**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Andreas Stihl AG & Co. KG, Fellbach

Grundlagen

Für eine Vielzahl von Unternehmen stellt das Geschäft mit Zubehör- und Ersatzteilen eine zentrale Einnahmequelle im After-Sales-Bereich dar. In Branchen wie dem Maschinenbau und der Automobilindustrie werden die Gewinne aus dem Aftermarket oft als höher erachtet als die Erlöse aus dem ursprünglichen Produktverkauf [4]. Für Unternehmen ergibt sich daraus die Herausforderung, die zukünftige Nachfrage nach Ersatz- und Zubehörteilen möglichst präzise vorherzusagen.

Die Zielsetzung dieser Arbeit besteht in der Entwicklung eines Konzepts zur datenbasierten Modellierung von Zubehör- und Ersatzteilverkäufen. Die vorliegende Untersuchung hat zum Ziel, auf Basis historischer Absatzdaten zu ermitteln, inwiefern sich praxistaugliche Prognosemodelle realisieren lassen. Der Fokus liegt dabei nicht ausschließlich auf der Erzielung einer möglichst hohen Prognosegenauigkeit, sondern auch auf der Interpretierbarkeit der Modellergebnisse. Letzteres ist insbesondere für den Einsatz in der Unternehmenspraxis von Relevanz, um nachvollziehbare und handlungsleitende Erkenntnisse aus den Modellen ableiten zu können.

Methodik

Zur Entwicklung eines datenbasierten Prognosemodells dient ein entworfener Analyseprozess Abb. 1 als methodische Leitlinie. Dieser strukturierte Ablauf unterstützt die systematische Konzeption, Umsetzung und Bewertung geeigneter Vorhersagemodelle und vereint bewährte Verfahren der Datenanalyse mit praxisrelevanten Anforderungen wie Modellinterpretierbarkeit und Optimierungspotenzial.

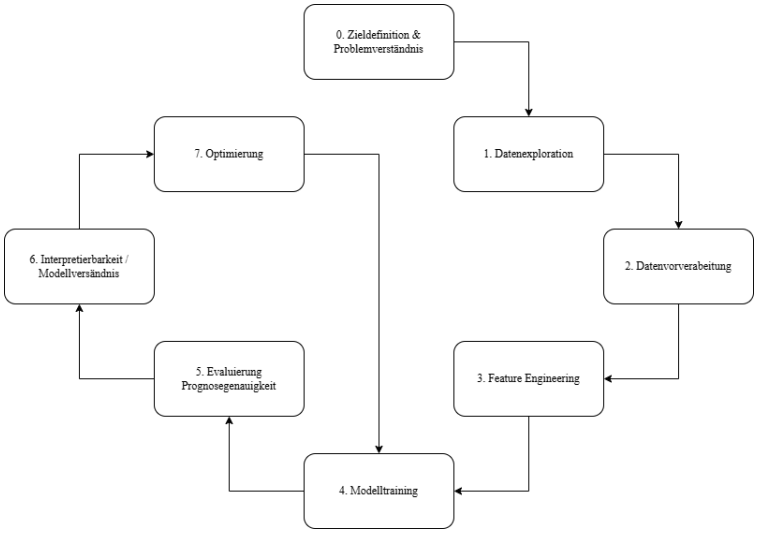


Abb. 1: Modellierungsprozess [2]

Auf Basis einer umfassenden Literaturrecherche zu praxisnahen Anwendungsfällen und Studien im Bereich der Zeitreihenprognose wurden verschiedene Modellierungsansätze analysiert.
Aufgrund ihrer signifikant hohen Prognoseleistung und ihrer breiten Anwendung in vergleichbaren Szenarien wurden drei Gradient-Boosting-Verfahren ausgewählt: XGBoost, LightGBM und CatBoost.
Alle drei Ansätze basieren auf dem Prinzip des Boosting, bei dem schwache Lernmodelle (i. d. R. Entscheidungsbäume) zu einem leistungsfähigen Gesamtmodell kombiniert werden.
XGBoost überzeugt durch eine hohe Rechenleistung und robuste Regularisierungstechniken. LightGBM setzt auf einen histogrammbasierten Trainingsansatz, während CatBoost mit Ordered Boosting gezielt kategorische Daten verarbeitet [5], [1], [3].

Vorläufige Ergebnisse

Im Rahmen eines ersten Modellierungsansatzes wurde ein Baseline-Modell unter Verwendung des XGBoost-Algorithmus entwickelt, um einen ersten Einblick in das prognostische Potenzial zu erhalten. Dieses initiale Modell diente der Validierung des methodischen Vorgehens und ermöglichte eine erste Einschätzung der Modellgüte unter realitätsnahen, aber bewusst vereinfachten Annahmen.
Die Analyse konzentrierte sich auf ein reduziertes Szenario, bei dem die Verkaufszahlen eines spezifischen Ersatzteils in direkter Abhängigkeit zum Absatz eines korrespondierenden Hauptprodukts betrachtet wurden. Dieses stark vereinfachte Setup erlaubt eine gezielte Modellierung eines klar erkennbaren Zusammenhangs und bildet eine gute Ausgangsbasis für weiterführende Modellierungsstufen.
Zur Erfassung zeitlicher Muster wurden sogenannte Lag-Features eingesetzt, die die Absatzmengen der vergangenen sieben Jahre als erklärende Variablen berücksichtigen. Diese Zeitverschiebungen sollen dabei helfen, wiederkehrende Strukturen, Saisonalitäten oder Nachfragetrends im Datenverlauf zu erkennen. Auf zusätzliches Feature Engineering wurde in dieser frühen Phase bewusst verzichtet, um die Modellkomplexität gering zu halten und mögliche Überanpassung zu vermeiden.
Die ersten Prognoseergebnisse, visualisiert in Abb. 2, zeigen, dass das Modell in der Lage ist, übergeordnete Trendverläufe zuverlässig zu identifizieren. Gleichzeitig wird deutlich, dass bei stark unregelmäßigen Absatzbewegungen die Genauigkeit der Vorhersage noch begrenzt ist. Dies deutet auf ein Verbesserungspotenzial im Hinblick auf die Modellstruktur sowie die Berücksichtigung zusätzlicher Einflussgrößen hin, die in zukünftigen Modellierungsiterationen einbezogen werden sollen.

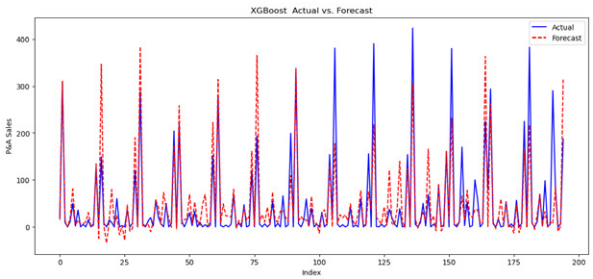


Abb. 2: P&A Sales Forecast – XGBoost [2]

Ausblick

Im weiteren Verlauf der Arbeit ist eine Weiterentwicklung des bestehenden Modellierungskonzepts vorgesehen, bei der das zugrunde liegende Prognoseszenario realistischer und praxisnäher gestaltet wird. Konkret soll das Modell künftig nicht nur auf die Vorhersage einzelner Ersatzteilverkäufe beschränkt bleiben, sondern um die gleichzeitige Abbildung mehrerer unterschiedlicher Ersatzteile erweitert werden. Auf diese Weise können komplexere Zusammenhänge innerhalb des Aftermarket-Geschäfts besser berücksichtigt und analysiert werden.
Darüber hinaus wird angestrebt, das bisherige Feature-Set, das hauptsächlich auf zeitlich verzögerten Variablen basiert, um weitere erklärende Einflussgrößen zu ergänzen. Dazu zählen beispielsweise externe Faktoren, produktspezifische Merkmale oder saisonale Effekte, die potenziell signifikanten Einfluss auf das Nachfrageverhalten haben.
Ein besonderes Augenmerk liegt dabei auf der Interpretierbarkeit der verwendeten Modelle. Neben der reinen Prognosegenauigkeit ist es entscheidend, dass die getroffenen Vorhersagen nachvollziehbar sind und konkrete Handlungsempfehlungen für die betriebliche Praxis ableitbar machen. Zu diesem Zweck sollen moderne Methoden der Modellinterpretation wie Feature Importance oder SHAP-Werte zum Einsatz kommen. Diese Verfahren ermöglichen es, den Einfluss einzelner Inputvariablen auf das Modellverhalten transparent darzustellen und so datengetriebene Entscheidungen gezielt zu unterstützen.

Literatur und Abbildungen

[1] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016.

[2] Eigene Darstellung.

[3] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Ya Liu. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 2017.

[4] Cohen Morris, Agrawal Narendra, and Agrawa Vipul. Winning in the aftermarket. *Harvard business review*, 2006.

[5] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Dorogush, and Andrey Gulin. CatBoost: Unbiased Boosting with Categorical Features. In *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc., 2018.

Künstliche Intelligenz im Qualitätsmanagement: Identifikation und Evaluierung relevanter Funktionen

Jordi Beeck **Andreas Rößler**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Roxtra GmbH, Göppingen

Einführung

Mit dem Aufschwung der künstlichen Intelligenz (KI) im Alltag werden auch Unternehmenssysteme überarbeitet und um neue Features erweitert. Durch die Fortschritte in der KI-Entwicklung lassen sich bestehende und neue Herausforderungen besser bewältigen. Im Bereich des Qualitätsmanagements (QM) ist das Potenzial der KI noch nicht ausgeschöpft. Unternehmen, die sich mit Qualitätsmanagement befassen, erkennen ebenso die Möglichkeiten der KI. Dabei muss QM-Software mit den Entwicklungen im KI-Bereich Schritt halten und Nutzern einen echten Mehrwert bieten. Gerade deshalb müssen potenzielle Features zunächst identifiziert und evaluiert werden, da Large Language Models (LLMs) nicht alle im Arbeitsalltag anfallenden Aufgaben bewältigen können. [2]

Ziel der Arbeit

Es sollen Features gefunden werden, die durch künstliche Intelligenz einen Mehrwert für Benutzer von Qualitätsmanagement-Software bieten. Dabei sollen nicht nur neue Features gefunden, sondern auch bestehende Features innerhalb des Systems durch KI erweitert werden. Die Evaluierung der Features erfolgt durch eine Bewertungsmatrix, welche wiederum als Basis für die vereinzelte Umsetzung von Proof of Concepts (PoCs) dient. Diese Umsetzung wird maßgeblich von der resultierenden Bewertung der Features, einer durchgeführten Kundenumfrage und den internen Zielen der Roxtra GmbH beeinflusst. Mit der Umsetzung der PoCs werden mögliche Features für das digitale Qualitätsmanagementsystem roXtra vorgestellt.

Large Language Models

Large Language Models sind neuronale Netze, die auf großen Datenmengen trainiert wurden. Diese Datenmenge ermöglicht LLMs das Interpretieren und Generieren von textbasierter Sprache. Dadurch lassen sich LLMs als generative künstliche Intelligenz einordnen. Mithilfe von Prompts, also gezielten Texteingaben, können LLMs Anweisungen zu verschiedenen Aufgaben erhalten und sind somit vielseitig einsetzbar. [4]

Retrieval Augmented Generation

Retrieval Augmented Generation (RAG) ist ein Verfahren, mit dem Large Language Models mit zusätzlichem Wissen versorgt werden können. Dabei werden nach Anfrage des Benutzers relevante Informationen ausgemacht und dem Kontext der KI zur Verfügung gestellt. Im Fall einer QM-Software werden relevante Dokumente durch eine semantische Suche ermittelt und der KI übergeben. Dies wird durch die vorangehende Indexierung der verfügbaren Dokumente ermöglicht. Dabei transformiert ein Embedding-LLM den Textinhalt der Dokumente in Vektoren, welche die semantische Bedeutung der Texte widerspiegeln. Mit Anfrage eines Benutzers werden passende Dokumente über die Vektordatenbank ausfindig gemacht. Im Anschluss erhält das LLM die Anfrage des Benutzers und den Textinhalt aller relevanten Dokumente, sodass eine Antwort generiert werden kann (siehe Abbildung 1). Durch die Bereitstellung von externem, relevantem Wissen können Antworten verbessert und Halluzinationen von LLMs vermindert werden. [3]

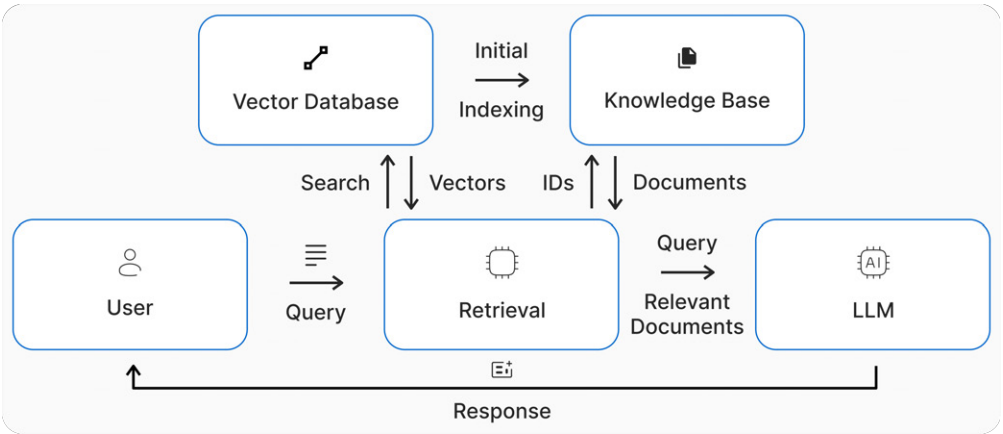


Abb. 1: Ablauf der Retrieval Augmented Generation [1]

Identifikation

Das Qualitätsmanagementsystem roXtra bietet verschiedene Module. Diese Module behandeln Dokumentenlenkung, Prozessautomatisierung, Risiken, Maßnahmen und Audits. Unter Berücksichtigung der bestehenden Funktionen in diesen Modulen und aktueller Potenziale der künstlichen Intelligenz wurden Verbesserungen und neue Features identifiziert. Wie diese Features auf technischer Ebene umgesetzt werden können, wird auf Grundlage der aktuellen Technologien und Trends der IT-Branche entschieden.

Bewertung

Die Bewertung der Features erfolgt über eine Bewertungsmatrix. Diese Bewertungsmatrix besteht aus zwei gleichgewichteten Bereichen, den Kriterien des Qualitätsmanagements und den technischen Kriterien. Innerhalb der beiden Bereiche werden Punkte auf Basis einzelner Kriterien vergeben. Es können insgesamt 100 Punkte erreicht werden, wobei die Bepunktung eines Kriteriums stufenweise in 5-Punkte-Schritten erfolgt. Die Bewertung wird durch eine Kundenumfrage unterstützt. Diese erfasst die Bedürfnisse der Kunden und identifiziert Trends und Stimmungen zum Thema KI aus der Wirtschaft.
Die Kriterien des Qualitätsmanagements lauten wie folgt:

- Qualitätssteigerung: Gibt an, inwiefern das Ergebnis eines Arbeitsvorgangs durch das neue Feature verbessert werden könnte. Es sind maximal 20 Punkte zu vergeben.
- Effizienzsteigerung: Bewertet, wie viel Zeit bei einem Arbeitsvorgang eingespart werden könnte. Es sind maximal 20 Punkte zu vergeben.
- Prozessintegration: Zeigt die Herausforderungen der Integration in bestehende Prozesse und Arbeitsabläufe. Es sind maximal 10 Punkte zu vergeben.

Die zweite Hälfte der Bewertungsmatrix bilden die technischen Kriterien:

- Umsetzbarkeit: Erfasst die technische Umsetzbarkeit unter Berücksichtigung des bestehenden Systems. Es sind maximal 15 Punkte zu vergeben.
- Kosten: Untersucht das relative Ausmaß der Kosten für Umsetzung und Betrieb. Es sind maximal 15 Punkte zu vergeben.
- Usability: Beurteilt die Gebrauchstauglichkeit, Bedienbarkeit und Effektivität. Auch die Relevanz aus Nutzersicht fließt ein. Es sind maximal 10 Punkte zu vergeben.
- Vielseitigkeit: Evaluiert die Konfigurier- und Erweiterbarkeit. Mögliche weitere Features werden auch betrachtet. Es sind maximal 10 Punkte zu vergeben.

Ausblick

Die identifizierten Funktionen werden durch die Bewertungsmatrix betrachtet und bewertet. Nach der Bewertung der Features erfolgen vereinzelte Umsetzungen als Proof of Concepts. Aufgetretene oder mögliche Herausforderungen werden aufgezeigt. Es entsteht eine Auflistung der Funktionen mit Potenzialen des Qualitätsmanagements, Herausforderungen der technischen Umsetzung und Implikationen für zukünftige Entwicklungen.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Fabrizio Dell'Acqua, Edward McFowland III, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani. Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. https://ssrn.com/abstract=4573321, 2023.

[3] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-Augmented Generation for Large Language Models: A Survey. https://arxiv.org/abs/2312.10997, 2024.

[4] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. https://arxiv.org/abs/2302.11382, 2023.

Konzeption und Umsetzung einer event-getriebenen Architektur für skalierbare Webanwendungen im Bereich Lerninhaltserstellung

Malte Budig **Jörg Nitzsche**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Savvi Learning GmbH, Kornwestheim

Einleitung & Motivation

Die Einbindung von Künstlicher Intelligenz (KI) in Webanwendungen eröffnet neue Funktionalitäten, bringt jedoch auch technische Herausforderungen mit sich. Insbesondere die langen Verarbeitungszeiten vieler KI-Modelle stehen im Widerspruch zu den Erwartungen moderner Nutzer an schnelle und reaktive Interfaces. Klassische synchrone Webarchitekturen stoßen hier an ihre Grenzen, da Anfragen häufig Timeout-Grenzen überschreiten und somit zu einer unterbrochenen Nutzererfahrung führen [4].
Gleichzeitig erwarten Nutzer in Echtzeitsystemen sofortige Rückmeldungen, etwa durch das sukzessive Anzeigen von Zwischenergebnissen während der KI-Verarbeitung. Asynchrone Kommunikationsverfahren wie WebSockets oder Server-Sent Events (SSE) bieten hierfür geeignete Lösungen, da sie eine fortlaufende Verbindung zwischen Server und Client ermöglichen und so die Benutzerfreundlichkeit deutlich verbessern. Vor diesem Hintergrund stellt sich die zentrale Forschungsfrage: **Wie kann eine skalierbare und wartbare Event-Driven Architecture in Webanwendungen eingesetzt werden, um asynchrone Kommunikation zu ermöglichen?**
Ziel dieser Arbeit ist es daher, eine modulare Webarchitektur zur Erstellung von KI-gestützten Softskill-Trainingsinhalten zu entwickeln, die auf einem Event-Driven-Ansatz basiert. Diese Architektur soll eine entkoppelte, asynchrone Kommunikation zwischen Frontend, Backend, Message-Broker und KI-Service ermöglichen. Durch den kontinuierlichen Informationsfluss sollen Nutzer jederzeit über den Fortschritt der KI-gestützten Inhaltserstellung informiert bleiben, was nicht nur die Interaktivität, sondern auch die Wart- und Skalierbarkeit des Systems verbessert.

Herausforderungen

Die Integration von KI in Webanwendungen bringt mehrere Herausforderungen mit sich. Erstens benötigen KI-Modelle signifikante Verarbeitungszeit, die nicht mit den Erwartungen an schnelle Webantworten übereinstimmt [1]. Zweitens können lange Wartezeiten während der KI-Verarbeitung zu Frustration führen, wenn nicht angemessen kommuniziert wird. Eine transparente und kontinuierliche Rückmeldung an den Nutzer ist daher essenziell. Drittens stellt die horizontale Skalierung von WebSocket- oder SSE-Verbindungen eine Herausforderung dar, da eine Verbindung typischerweise nur zwischen einem Server und dem Browser besteht. Eingehende und ausgehende Informationen müssen daher effizient an alle Instanzen verteilt werden, um eine konsistente Kommunikation sicherzustellen [5].

Event-Driven Architecture (EDA)

Zur Bewältigung der genannten Herausforderungen wird eine Event-Driven Architecture (EDA) implementiert, die folgende Vorteile bietet: Durch die Entkopplung der Systemkomponenten können diese unabhängig voneinander entwickelt, skaliert und gewartet werden. EDA ermöglicht eine effiziente Verarbeitung von Ereignissen, wodurch langlaufende Prozesse wie die KI-gestützte Inhaltserzeugung den Hauptanwendungsfluss nicht blockieren. Die Architektur unterstützt die horizontale Skalierung durch den Einsatz von Message Brokern, die eingehende Events speichern und gezielt an die zuständigen Backend-Komponenten verteilen [2].

Komponenten der Architektur

Die Anwendung umfasst vier Hauptkomponenten:
Das Frontend stellt die Benutzeroberfläche bereit und verarbeitet Nutzeraktionen. Es empfängt kon-

tinuierlich Updates vom Server über Server-Sent Events (SSE) und sorgt für Wiederverbindungen bei Unterbrechungen. Das Backend ist verantwortlich für das Event-Handling, die API-Kommunikation und die Weiterleitung von Ereignissen an das Frontend. Es ist modular aufgebaut und kann unter Last auf mehrere Instanzen verteilt werden. Der Message Store fungiert als zentrales Element der Entkopplung, speichert eingehende Events und verteilt diese gezielt an die zuständigen Backend-Komponenten. Der AI Generation Service übernimmt die Generierung der Inhalte auf Basis von Benutzereingaben, verarbeitet Anfragen asynchron, speichert Zwischenergebnisse und sendet entsprechende Events an den Message Store.
Durch diese Architektur wird eine reaktive Benutzeroberfläche gewährleistet, die den Nutzer kontinuierlich über den Fortschritt informiert, während im Hintergrund Inhalte generiert werden. Zudem ermöglicht die lose Kopplung eine einfache Erweiterung und Wartung des Systems.

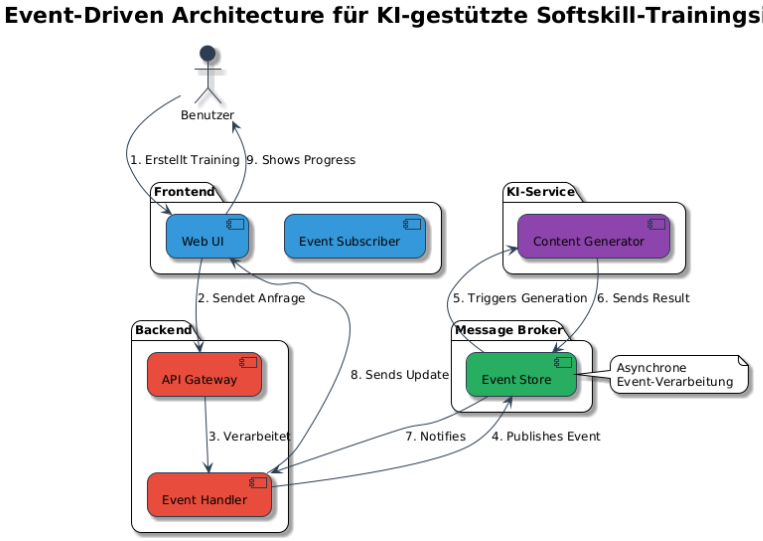


Abb. 1: Event-Driven Architecture für KI-gestützte Softskill-Trainingsinhalte [3]

Dieses Diagramm visualisiert die Interaktion zwischen den Komponenten in einer Event-Driven Architecture für KI-gestützte Softskill-Trainingsinhalte. Der Benutzer interagiert mit der Weboberfläche, die Anfragen an das Backend sendet. Das Backend verarbeitet die Anfragen und veröffentlicht Ereignisse an den Message Broker (z. B. Kafka oder Redis). Der KI-Service konsumiert diese Ereignisse, generiert Inhalte und sendet die Ergebnisse zurück über den Message Broker an das Backend, welches die Weboberfläche aktualisiert [2].

Ausblick: Implementationen und Technologien

Die in dieser Arbeit vorgestellte Architektur wurde mit modernen und bewährten Technologien umgesetzt, die eine hohe Skalierbarkeit, Wartbarkeit und Flexibilität gewährleisten. Für das Frontend und Backend kommt Next.js zum Einsatz, das sowohl serverseitiges Rendering als auch eine effiziente API-Entwicklung ermöglicht. Die Wahl von Next.js erlaubt es, eine konsistente Codebasis für beide Schichten zu nutzen und so Entwicklungsaufwand zu reduzieren sowie die Wartung zu vereinfachen.
Als Message Store wird Redis mit Pub/Sub-Mechanismus verwendet. Redis bietet eine performante und leichtgewichtige Lösung für die asynchrone Event-Verteilung zwischen den Systemkomponenten. Durch die Nutzung von Pub/Sub können Events effizient an mehrere Backend-Instanzen verteilt werden, was insbesondere für die horizontale Skalierung und die Realisierung von Echtzeit-Updates im Frontend entscheidend ist [5].
Der KI-Service ist als externe API angebunden. Diese Entkopplung ermöglicht es, verschiedene KI-Modelle oder Anbieter flexibel zu integrieren, ohne die Kernarchitektur der Anwendung anpassen zu müssen. Die Kommunikation mit dem KI-Service erfolgt asynchron, sodass längere Verarbeitungszeiten die Nutzererfahrung nicht beeinträchtigen.

Literatur und Abbildungen

[1] Işıl Karabey Aksakallı, Turgay Çelik, Ahmet Burak Can, and Bedir Tekinerdogan. Deployment and communication patterns in microservice architectures : A systematic literature review. *The Journal of Systems & Software*, 2021.

[2] Ashwin Chavan. Exploring event-driven architecture in microservices- patterns, pitfalls and best practices. *International Journal of Science and Research Archive*, 2021.

[3] Eigene Darstellung.

[4] Alok Dubey. Enhancing Real Time Communication and Efficiency With Websocket. *International Research Journal of Engineering and Technology*, 10, 2023.

[5] Vaibhav Vudayagiri. SCALABLE AI-DRIVEN MICROSERVICES ARCHITECTURES FOR DISTRIBUTED CLOUD ENVIRONMENTS. *International Journal of Computer Engineering and Technology*, 2024.

Radarbasierte Konturenerkennung durch maschinelles Lernen mit LiDAR-Referenzdaten

Luca Cais **Clemens Klöck**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung:

Radarsysteme nehmen in technischen Anwendungen eine zunehmend wichtigere Rolle ein. Sie ermöglichen die Erfassung relevanter Informationen zur Umgebung wie Entfernungen, Geschwindigkeiten, Winkelpositionen sowie – in begrenztem Umfang – Rückschlüsse auf das Material reflektierender Objekte. Im Automotive-Bereich wird diese vielseitige Technologie bereits eingesetzt. Sie ist essenziell für den Spurwechselassistenten, die Einpark-hilfe und die Objekterkennung [4]. Obwohl diese Technologie so vielseitig ist und viele Anwendungsbereiche besitzt, stößt sie in bestimmten Bereichen an ihre Grenzen.

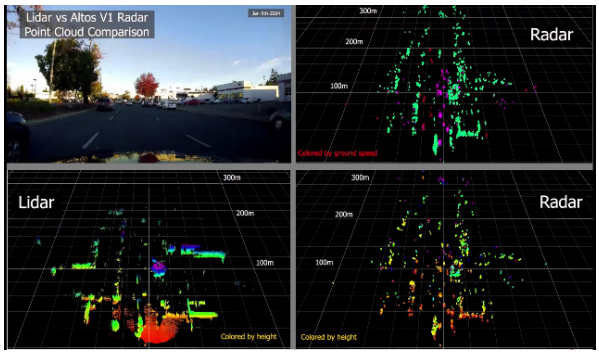


Abb. 1: Vergleich von Radar- und Lidarpunktwolken [1]

Die präzise Erfassung von Objektkonturen ist mit Radarsystemen allein nur eingeschränkt möglich. Dahingegen ermöglichen Lidarsysteme eine präzise Messung von Objektkonturen im Raum. Aus diesen Gründen werden Lidarsysteme ebenfalls im Automotive-Bereich verwendet. Lidarsysteme besitzen jedoch den entscheidenden Nachteil, dass sie wetterabhängig sind. Bei Regen oder starkem Rauch nimmt die Präzision der Messungen enorm ab.

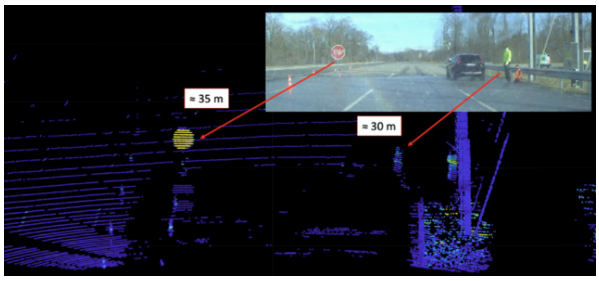


Abb. 2: Lidar Punktwolke bei Regen [3]

Radarsysteme sind dahingegen wetterunabhängig. Daher ist es von großer Bedeutung zu untersuchen, wie genau Radarsysteme Objekte erkennen und welche Informationen dabei zuverlässig extrahiert werden können. Ein besseres Verständnis der Radarsignalverarbeitung könnte dazu beitragen, Radarsysteme gezielt so zu trainieren, dass sie Objekte annähernd so präzise wie Lidarsysteme identifizieren können. Dies würde das Potenzial wetterunabhängiger Objekterkennung im Automotive-Bereich erheblich steigern.

Zielsetzung:

Das Ziel dieser Bachelorarbeit ist es, zu untersuchen, wie ein Radar die Konturen im Vergleich zu einem Lidar wahrnimmt. Dazu soll ein echtzeitfähiges Testsystem entwickelt werden, das LIDAR- und Radardaten in einem Raum misst, Objekte erfasst und in einem gemeinsamen Koordinatensystem darstellen kann. Das Ziel dieser gemeinsamen Darstellung ist es, Datensätze zu sammeln, die zum Trainieren eines neuronalen Netzes verwendet werden können. Die LIDAR-Daten sollen hierbei als ground truth verwendet werden. Das neuronale Netz soll in der Lage sein, aus den ungenauen Radarmessungen präzise Informationen zu rekonstruieren, wie sie ein LIDAR-System liefern würde. Über diesen Umweg wäre es möglich, Radardaten präziser zu erfassen und Objekte dadurch genau zu bestimmen.

Vorgehensweise:

Zur Umsetzung des Projekts wird eine 3-Teilige Aufgabenstruktur verfolgt. Der erste Schritt des Projekts besteht in der Inbetriebnahme und Konfiguration beider Sensorsysteme. Dabei wird zunächst der Lidar Pandar64 aufgesetzt und kalibriert. Parallel dazu erfolgte die Einrichtung des Radars AWR1642boost. Ein Zentraler Bestandteil des Projekts besteht in der räumlichen Abstimmung beider Sensoren aufeinander. Um Datensätze beider Sensoren miteinander vergleichen zu können müssen ihre Positionen zueinander exakt bestimmt werden. Hierzu benötigt man die Positionen der beiden Systeme zueinander im Raum wie auch die translationalen und rotatorischen Lageparameter. Zur Umsetzung wird ein Python-Algorithmus genutzt, der die Datensätze beider Sensoren einliest und sie anschließend in ein gemeinsames Koordinatensystem transformiert. Nach erfolgreicher Synchronisation und Transformation kann mit der Datenerfassung begonnen werden. In einem Testraum werden hierfür verschiedene Objekte positioniert und simultan mit beiden Sensoren aufgezeichnet. Ziel ist es hierbei die gesammelten Lidardaten als verlässliche Referenz zu verwenden, um daraus Trainingsdaten für das neurale Netz zu erzeugen. In einem nächsten Schritt folgt dann die Aufbereitung und Vorverarbeitung der gesammelten Daten durch Segmentierung und Labeling. Diese vorbereiteten Daten bilden die Grundlage für die Entwicklung und das Training des neuronalen Netzes, welches später allein mit Radardaten zuverlässig Objekte erkennen soll.

Einsatzzweck:

Radarbasierte Konturenerkennung findet in vielen verschiedenen Bereichen ein Nutzen. Überall dort, wo herkömmliche Sensoren durch Umgebungsbedingungen wie Regen, Hitze oder Nebel an ihre Grenzen stoßen, liefert der Radar zuverlässig Ergebnisse [2]. Ein konkretes Anwendungsbeispiel ist die Integration eines solchen Systems in die Ausrüstung von Feuerwehrmännern. In verrauchten oder brennenden Räumen ist es für Einsatzkräfte oft schwierig oder unmöglich, sich einen Überblick zu verschaffen. Das System kann dabei unterstützen, Objekte oder Personen im Raum zu erkennen und diese Informationen an den Feuerwehrmann in Echtzeit weitergeben.

Literatur und Abbildungen

[1] Radar Altos. Altos Radar & Lidar point cloud comparison [Video]. https://www.youtube.com/watch?v=ymeRYplU8A, 2024.

[2] Jens Klare. Bildgebende Radarsensorik. In *ITGnews*, pages 14–16. ITG, 2018.

[3] TTLab Mechlab. LiDAR im Regen // LiDAR@Rain – Mechlab. https://mechlab.de/lidar-im-regen-lidarrain/, 2024.

[4] Konrad Reif. *Automobilelektronik Eine Einführung für Ingenieure 4*. Vieweg +Teubner Verlag, 2012.

Architektur und prototypische Implementierung einer Runtime für die Anbindung unterschiedlicher Scriptsprachen an eine SPS am Beispiel von Python mit einer REST-API für Konfiguration und Statusanzeige

Selim Cetin **Rainer Keller**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Robert Bosch GmbH, Stuttgart-Feuerbach

Abstract

Bosch Connected Industry (BCI) bietet Lösungen im Bereich der Maschinenprogrammierung an, darunter auch Anwendungen für industrielle Bildverarbeitung (Machine Vision). Ein zentraler Bestandteil dieser Lösungen ist die Objekterkennung mittels Kameras, beispielsweise zur Zählung von Werkstücken. Die technologische Grundlage bildet das Bildverarbeitungs-Framework HALCON der Firma MVTec. Die Implementierung der Bildverarbeitungslogik erfolgt in der proprietären HALCON-Skriptsprache innerhalb der Entwicklungsumgebung HDevelop. Diese Machine-Vision-Lösung wird in einem Umfeld mit speicherprogrammierbaren Steuerungen (SPS bzw. PLC) und Computern eingesetzt. Die Bildverarbeitungssysteme, beispielsweise ein Windows-Computer, kommunizieren über OPC UA mit der SPS.

Einleitung

Zur Steuerung der Objekterkennung und Bildverarbeitung wird eine Runtime benötigt, die eine OPC-UA-Verbindung zur SPS aufbaut. Die derzeit eingesetzte Runtime basiert auf C# .NET und wird unter dem Namen HDevEngine.Exe (HDE2) geführt. HDE2 empfängt Steuerbefehle von der SPS – beispielsweise zum Starten, Zurücksetzen oder Kalibrieren – und führt daraufhin HALCON-Skripte aus. Diese interagieren mit den angeschlossenen Kameras und übernehmen die Bildverarbeitung. Die Runtime fungiert als zentrale Steuerungseinheit der HALCON-Umgebung: Sie initialisiert das System, startet eine Endlosschleife zur Abarbeitung eingehender Befehle und deinitialisiert es beim Empfang eines Shutdown-Befehls. Innerhalb dieser Schleife werden benutzerdefinierte Skripte ausgeführt, die zuvor mit HDevelop erstellt wurden.

Problemstellung

Ziel ist es, zusätzlich zur bisherigen Bildverarbeitung mit HALCON auch die Nutzung von Python für Aufgaben außerhalb der Bildverarbeitung zu ermöglichen. In der bestehenden Architektur würde dies einen tiefgreifenden Umbau erfordern und mit erheblichem redundantem Code einhergehen – insbesondere für wiederkehrende Abläufe wie die Kommunikationslogik oder die Steuerungsschleife. Darüber hinaus besteht der Wunsch, den Status der Runtime über eine REST-API abrufen zu können.

Zielsetzung

Ziel der Bachelorarbeit ist die Entwicklung einer modularen Code-Architektur mithilfe von C#-Interfaces bzw. abstrakten Klassen, bei der der Bildverarbeitungsprozess als generischer „DataProcessor“ definiert wird. Auf dieser Grundlage können konkrete Implementierungen für Python und HALCON erfolgen, die bei Bedarf per Dependency Injection ausgetauscht werden. Dies ermöglicht es dem Benutzer, flexibel zwischen Python- und HALCON-basierten Anwendungen zu wählen – ohne die zugrunde liegende Systemarchitektur anpassen zu müssen.

Motivation und Nutzen

HALCON ist hochgradig spezialisiert auf die industrielle Bildverarbeitung. Python bietet als weltweit verbreitete Universalprogrammiersprache sowohl Bibliotheken für verschiedenste Anwendungsgebiete als auch Zugriff auf einen großen Entwicklerpool. Die Unterstützung beider Technologien in einer gemeinsamen Automatisierungslogik verringert den Inbetriebnahmeaufwand, erleichtert die Wartung und spart so Zeit und Kosten.

Technologischer Hintergrund: Python

Python ist eine der am weitesten verbreiteten Programmiersprachen. Sie unterstützt sowohl objektorientierte als auch funktionale Programmierparadigmen. Durch die dynamische Typisierung fällt der Einstieg in die Sprache leicht, da typische Kompilierfehler statisch typisierter Sprachen vermieden werden. Entgegen der landläufigen Meinung ist Python nicht rein interpretiert: Der Quellcode (.py) wird zunächst in einen Abstract Syntax Tree (AST) geparst und anschließend in Bytecode (.pyc) kompiliert. Dieser wird schließlich vom Python-Interpreter zeilenweise ausgeführt. Die am weitesten verbreitete Implementierung ist CPython, die offizielle Referenzimplementierung der Sprache, entwickelt in C. Wenn von „Python“ die Rede ist, ist in der Regel CPython gemeint.

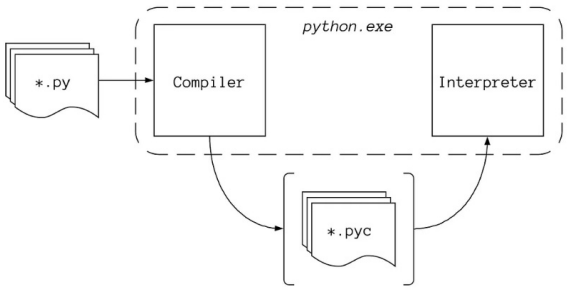


Abb. 1: High-Level Überblick des Python Interpreters [2]

Implementierung

Im ersten Schritt wird ein C#-Interface definiert, das die Anforderungen an einen „DataProcessor“ beschreibt. Basierend darauf werden konkrete Implementierungen für HALCON, Python und ggf. weitere Technologien entwickelt. Als zentrale Steuerungseinheit dient eine ASP.NET-Core-Anwendung, die als Runtime fungiert. Diese enthält die Endlosschleife zur Befehlsverarbeitung sowie eine REST-API zur Statusabfrage. Für die Python-basierte Anwendung werden separate Skripte entwickelt, die sowohl die OPC-UA-Verbindung als auch die Befehlsverarbeitung übernehmen. Die Verbindung zur SPS erfolgt ausschließlich innerhalb der Python-Implementierung des DataProcessors. Dabei wird der gewünschte Steuerbefehl durch regelmäßiges „Polling“ eines spezifischen OPC-UA-Knotens abgefragt – die SPS sendet also keinen Befehl aktiv, sondern stellt diesen zum Abruf bereit.
Die konkrete Implementierung der jeweiligen Anwendungsalgorithmen ist nicht Bestandteil dieser Bachelorarbeit.

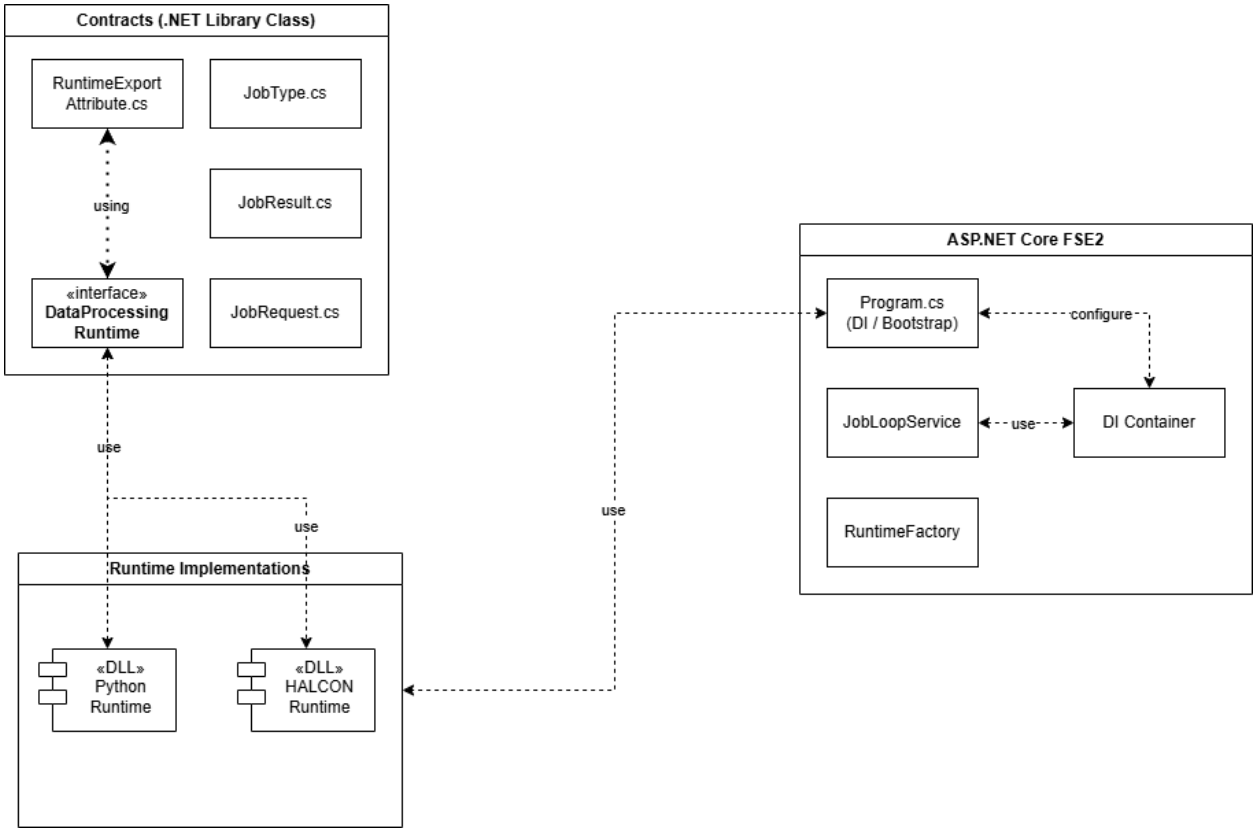


Abb. 2: Überblick Softwarearchitektur [1]

Fazit

Die im Rahmen dieser Bachelorarbeit entwickelte Architektur ermöglicht eine flexible und zukunftssichere Gestaltung industrieller Bild- und Datenverarbeitungssysteme. Durch die Entkopplung der Verarbeitungslogik von der zentralen Steuerungseinheit über generische Schnittstellen können verschiedene Skriptsprachen verwendet werden – ohne grundlegende Änderungen an der Systemlogik vornehmen zu müssen. Dies reduziert den Entwicklungsaufwand und vermeidet Redundanzen.

Die Unterstützung von Python und potenziell weiteren Skriptsprachen vergrößert den Anwendungsbereich und ermöglicht den Zugang zu einem breiten Entwicklerpool. Gleichzeitig bleibt die bestehende HALCON-Infrastruktur weiterhin nutzbar.

Insgesamt leistet die Arbeit einen Beitrag zur technologischen Offenheit und Modularisierung in der industriellen Automatisierung und zeigt praxisnah, wie weitere Anwendungsbereiche mit modernen Softwarekonzepten kombiniert werden können.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Michael Prantl. Python Internals: An Introduction. https://blog.sourcerer.io/python-internals-an-introduction-d14f9f70e583, 08 2020.

Evaluation von Wi-Fi 7 für die Industrielle Automatisierung

Simon Claus **Michael Scharf**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

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

Einleitung

Mit den wachsenden Anforderungen an drahtlose Netzwerke stoßen die aktuellen WLAN-Technologien an ihre Grenzen. Viele Bereiche wie Augmented Reality (AR), Virtual Reality (VR), das Streaming von 4k- und 8k-Videos und Cloud Gaming benötigen eine sehr hohe Datenrate und niedrige Latenz. Um darauf eine Antwort zu haben, entwickelt die IEEE 802.11-Arbeitsgruppe den neuesten Standard Wi-Fi 7. Damit soll vor allem die hohe Anforderung an die Datenrate erfüllt werden. In Bezug auf industrielle Automatisierung und „Industrie 4.0“ stellt sich die Frage, ab wann drahtlose Netzwerke geeignet sind und ob dieser Punkt bereits mit Wi-Fi 7 erreicht ist.

Problemstellung und Zielsetzung

Im Rahmen der Arbeit wird evaluiert, ob Wi-Fi 7 geeignet für die industrielle Automatisierung ist. Da Wi-Fi 7 sehr aktuell ist, gibt es eine große Varianz zwischen der Theorie und der praktischen Umsetzung in der Hardware und in der Treiberschicht. Um dies zu erforschen, werden verschiedene Wi-Fi 7-Module verglichen und auf unterstützte Funktionen überprüft. Darüber hinaus werden Latenz- und Bandbreitenmessungen durchgeführt, um das Potenzial der Module zu ermitteln und erste Aussagen über die Praxistauglichkeit zu treffen.
Abschließend werden mithilfe der Ergebnisse geeignete Anwendungsfälle in Bezug auf industrielle Automatisierung ermittelt und prototypisch umgesetzt. Diese Demonstration soll den Einsatz veranschaulichen und ein abschließendes Fazit liefern.

Wireless Fidelity 7 (Wi-Fi 7)

Wi-Fi 7, das auch unter dem Begriff „Extremely High Throughput (EHT)“ bekannt ist, ist der neueste WLAN-Standard, entwickelt von der IEEE 802.11-Arbeitsgruppe und soll unter idealen Bedingungen bis zu 30 Gbit/s gewährleisten. [4]
Dazu definierte die Arbeitsgruppe viele Funktionen, um höhere Bandbreite und geringere Latenz zu ermöglichen. Im Folgenden wird ein Teil der prominenten Funktionen aus dem Standard vorgestellt:
Multi-Link Operation (MLO): In einem bisherigen Netzwerk ohne Wi-Fi 7 können sich Geräte mit einem der drei Frequenzbänder 2,4 GHz, 5 GHz oder 6 GHz verbinden, um Daten auszutauschen. MLO ermöglicht die gleichzeitige Nutzung von zwei Frequenzbändern, was bedeutet, dass die Verbindung stabiler ist, eine höhere Datenrate möglich ist und somit auch eine geringere Latenz erreicht werden kann. [1]
4096 Quadratur-Amplitudenmodulation (4K-QAM): QAM beschreibt, wie viele Daten in einem Symbol codiert werden können. Das bisherige 1024-QAM kann 10 Bits codieren, 4096-QAM steigert dies auf 12 Bits, damit kann eine bis zu 20 % höhere Datenrate erzielt werden. Diese Funktion hat den Nachteil, dass ein hohes „Signal-to-Noise Ratio“ (SNR) erforderlich ist, was zu höheren Kosten und höherer Komplexität der Hardware führt. [3] SNR ist ein Messwert, der das Signal mit dem Hintergrundrauschen vergleicht. Ein hoher Wert bedeutet ein klares Signal mit wenig Hintergrundrauschen.
320MHz-Bandbreite: Je nach Frequenzband können Kanäle unterschiedlich große Bandbreiten haben, diese reichen von 20 MHz bis zu 160 MHz. Mit Wi-Fi 7 wird ein größere 320MHz-Bandbreite eingeführt. Diese kann nur im 6GHz-Frequenzband verwendet werden und verdoppelt den maximalen Nenndurchsatz im Vergleich zur 160MHz-Bandbreite. [4] Mit einer größeren Bandbreite ist es wahrscheinlicher, dass sich mehrere Nutzer überschneiden und stören. Um den Einfluss anderer zu verringern, wird die Funktion „Preamble Puncturing“ eingeführt. Die Funktion erlaubt es, störende Abschnitte auszublenden und damit einen Rückgang zu einer geringeren Bandbreite zu verhindern.

Analyse der Hardware

Zum Analysieren der Wi-Fi 7-Unterstützung in der Hardware werden die folgenden zwei Module untersucht: QCNCM865 von Qualcomm Technologies und MT7925 von MediaTek. Die Module werden mit dem

Betriebssystem Linux und mit dem „wireless-next“-Kernel getestet. „wireless-next“ ist ein Linux Kernel, der sehr nah am aktuellsten Entwicklungsstand ist.
Bei der Analyse der Qualcomm-Karte lässt sich schnell feststellen, dass der Treiber „ath12k“ instabil und die derzeit aktuelle Firmware defekt ist. Dies zeigen Fehlermeldungen im Kernel und ein nicht korrekt funktionierendes 6GHz-Frequenzband. Ohne das Frequenzband ist das Modul nicht Wi-Fi 7 und genau genommen auch nicht Wi-Fi 6E fähig. Andere Funktionen wie 320MHz-Bandbreite, 4K-QAM und MLO werden theoretisch unterstützt, sind aber durch die Treiber Probleme nicht benutzbar.
Die MediaTek-Karte hingegen funktioniert deutlich besser. Der Treiber „mt7925e“ liefert keine Fehlermeldungen und es können alle drei Frequenzbänder benutzt werden. Die Analyse zeigt, dass die optionalen Funktionen 320MHz-Bandbreite und 4K-QAM nicht unterstützt werden. Zusätzlich funktioniert MLO nicht, obwohl die Funktion verpflichtend für den Standard ist, was auf fehlende Treiberunterstützung hinweist.
Aktuell sind beide Module sehr mangelhaft in der Wi-Fi 7-Unterstützung. Es ist zu erwarten, dass die Probleme im Laufe der Zeit gelöst werden, da bei beiden Treibern viel entwickelt wird.

Benchmarking

Um das aktuelle Potenzial der beiden Karten zu ermitteln, werden Latenz- und Bandbreitenmessungen durchgeführt. Für die Messung wird ein einfaches Netzwerk, das in Abbildung 1 zu sehen ist, erstellt. Das Netzwerk besteht aus Client, Server, Access Point, Switch und Controller. Der Controller konfiguriert das WLAN, das vom Access Point bereitgestellt wird. Dabei können Parameter wie aktive Frequenzbänder, verwendete Wi-Fi-Standards und Funktionen wie MLO eingestellt werden. Der Client sowie der Server sind beides PCs auf denen Linux und die erforderliche Software installiert sind. Zudem wird beim Client eines der beiden Wi-Fi 7-Module eingebaut. Der Switch verbindet alle drahtgebundenen Komponenten. Die Route für das Benchmarking startet mit dem Client, dieser ist drahtlos mit dem Access Point verbunden und hat danach eine drahtgebundene Kommunikation zum Ziel (Server).

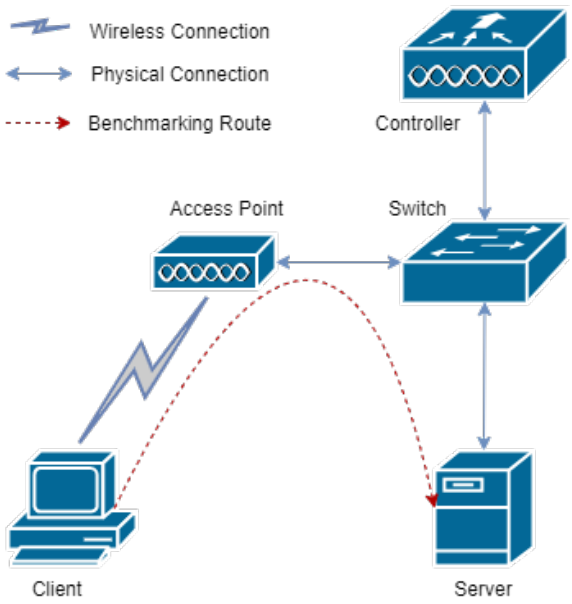


Abb. 1: Benchmarking-Netzwerk [2]

Die Bandbreitenmessungen werden mit dem Tool „iPerf3“ durchgeführt, welches das Testen der maximalen Bandbreite ermöglicht. Das Ergebnis mit der MediaTek-Karte beträgt eine maximale Bandbreite von 2,3 Gbit/s und mit der Qualcomm-Karte 1 Gbit/s. Beide Resultate sind weit weg von den versprochenen 30 Gbit/s, aber mit Treiber Problemen und fehlenden Funktionen zu erwarten.
Die Latenzmessungen, die in Abbildung 2 zu sehen sind, zeigen große Unterschiede zwischen beiden Karten. Das Diagramm zeigt die Häufigkeit pro Latenzbereich, der 0,5 ms groß ist, an. Ausreißer, die eine Latenz größer 30 ms haben, werden in einem Bereich zusammengefasst. Bei MediaTek ist ein hoher Jitter zu erkennen, die Werte sind von 2 ms bis 20 ms verstreut und es ist ein Wellenmuster zu sehen. Dieses Wellenmuster hat lokale Hochpunkte bei 4 ms, 6 ms, 13 ms und 18 ms, was bei einer Wi-Fi-Karte nicht vorkommen sollte. Bei Qualcomm hingegen sind alle Werte unter 10 ms (mit Ausnahme von Ausreißern), mit den meisten zwischen 3 ms und 4 ms, was ein deutlich geringerer Jitter ist.

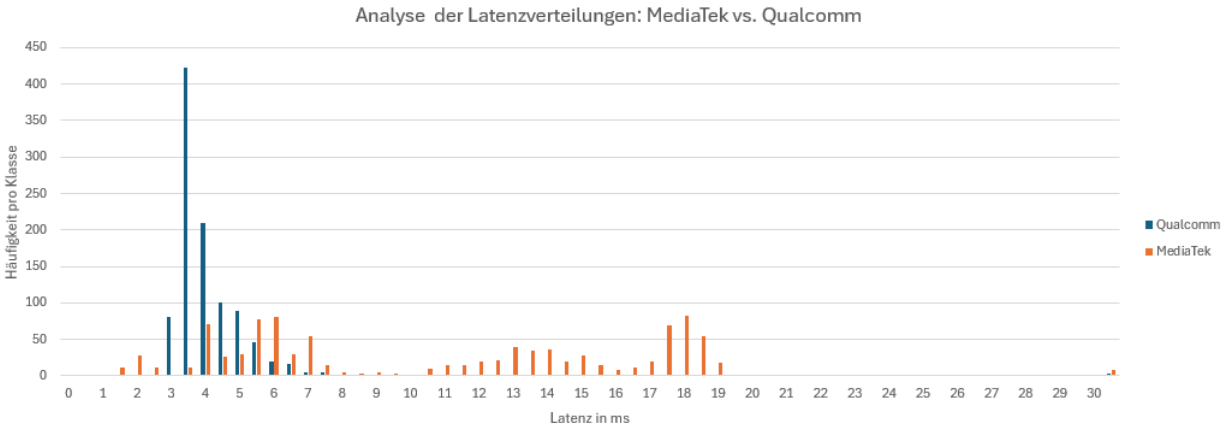


Abb. 2: Latenzmessung von MediaTek und Qualcomm über 1000 Messpunkte [2]

Dieses Ergebnis ist überraschend, da die Analyse suggeriert, dass die MediaTek-Karte mit mehr Wi-Fi 7-Funktionen bessere Ergebnisse liefern sollte als die Qualcomm-Karte.
Ein Jitter von 20 ms beim MediaTek-Modul ist für die Industrie nicht akzeptabel. Die Qualcomm-Karte ist mit geringerem Jitter für viele Use-Cases besser geeignet, allerdings ist die geringere Bandbreite ein Problem.

Ergebnis und Ausblick

Die Analyse zeigt, dass die Treiber noch viel Arbeit benötigen und viele Funktionen nicht unterstützt werden. Die Messungen zeigen, dass beide Karten die Bandbreite von 30 Gbit/s bei weitem nicht erreichen und die MediaTek-Karte ein großes Problem mit Jitter hat. Das Fazit ist, dass Wi-Fi 7 mit der aktuellen Treibersituation nicht für die industrielle Automatisierung geeignet ist.
Für die Zukunft ist zu erwarten, dass die Treiber weiterentwickelt und die genannten Probleme gelöst werden, um das Potenzial von Wi-Fi 7 voll auszuschöpfen.

Literatur und Abbildungen

[1] Shubhodeep Adhikari and Sindhu Verma. Analysis of Multilink in IEEE 802.11be. In *IEEE Communications Standards Magazine*, pages 52–58. IEEE, 6 edition, 2022.

[2] Eigene Darstellung.

[3] Cailian Deng, Xuming Fang, Xiao Han, et al. IEEE 802.11be Wi-Fi 7: New Challenges and Opportunities. In *IEEE Communications Surveys & Tutorials*, pages 2136–2166. IEEE, 22 edition, 2020.

[4] Xiaoqian Liu, Yuhan Dong, Yiqing Li, et al. IEEE 802.11be Wi-Fi 7: Feature Summary and Performance Evaluation. https://arxiv.org/abs/2309.15951, 07 2024.

GPS-Ortung, Cloud-Datenanbindung und Solarbetrieb zur Unterstützung der intelligenten Parkraumüberwachung

Mohamad Damen **Clemens Klöck**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

Die zunehmende Verdichtung städtischer Räume stellt moderne Infrastrukturen vor große Herausforderungen – insbesondere im Bereich der Mobilität und Parkplatzverfügbarkeit. Die Parkplatzsuche nimmt in vielen Städten einen erheblichen Teil der Verkehrszeit in Anspruch, was nicht nur zu Frustration bei Autofahrern, sondern auch zu erhöhtem CO -Ausstoß und unnötigem Verkehrsaufkommen führt. Technologien zur intelligenten Parkraumüberwachung bieten das Potenzial, diese Problematik effizient zu adressieren. Kamerabasierte Systeme, die freie und belegte Parkplätze automatisch erkennen, können dabei helfen, den innerstädtischen Verkehr zu entlasten und die Ressourcennutzung zu optimieren.Das bestehende Parkraumüberwachungssystem, das als Grundlage dieser Arbeit dient, basiert bereits auf lokaler Objekterkennung mittels Kamera. Es fehlt jedoch an drei zentralen Funktionen, die für einen mobilen, vernetzten und energieeffizienten Einsatz erforderlich sind: eine präzise Standorterfassung, eine cloudbasierte Speicherung der erfassten Daten sowie eine netzunabhängige Energieversorgung. Diese Arbeit beschäftigt sich mit der funktionalen Erweiterung dieses Systems durch genau diese Komponenten und schafft damit die Basis für eine flexiblere und skalierbare Lösung im Kontext moderner, intelligenter Parkraumanwendungen.

Ziel des Projekts

Ziel dieser Arbeit ist die funktionale Erweiterung eines bestehenden, kamerabasierten Parkraumüberwachungssystems. Dazu werden drei zentrale Komponenten integriert: ~~* ~ **GPS-Ortung:** um die Position der Überwachungseinheit exakt zu bestimmen und erkannte Parkplätze geografisch zuzuordnen. ~ **eine Cloud-Datenanbindung:** um Parkstatus, Zeitstempel und Positionsdaten zentral zu speichern und externen Anwendungen bereitzustellen. ~ **eine solargestützte Energieversorgung:** um das System unabhängig vom Stromnetz mobil und energieautark betreiben zu können. ~~~ Die Umsetzung erfolgt auf Basis eines Raspberry-Pi-Systems mit Kamera und GPS-Modul. Die Arbeit umfasst die Auswahl geeigneter Komponenten, deren Integration in das bestehende System, die Entwicklung der Software zur Datenerfassung und -übertragung sowie die praktische Erprobung des Gesamtsystems im Außenbereich.

Reverse Geocoding

Reverse Geocoding ist ein Verfahren, bei dem geografische Koordinaten – also Breitengrad (Latitude) und Längengrad (Longitude) – verwendet werden, um daraus eine menschenlesbare Adresse zu erzeugen. Diese Methode wird in vielen ortsbezogenen Anwendungen eingesetzt, um Positionen nicht nur technisch, sondern auch verständlich für Nutzer darzustellen. Im Rahmen dieser Arbeit wird der Dienst Nominatim, basierend auf den Daten von OpenStreetMap (OSM), für Reverse Geocoding verwendet. Nominatim sucht dabei nach dem nächstgelegenen passenden Objekt im OSM-Datenbestand, etwa einer Straße, einem Gebäude oder einem Stadtteil, und gibt daraus die zugehörige Adresse zurück. Dabei wird nicht zwangsläufig die exakte Adresse am Koordinatenpunkt gefunden, sondern diejenige, die dem Punkt räumlich und semantisch am nächsten ist. Die Abfrage erfolgt in der Regel über eine HTTP-Schnittstelle, bei der die Koordinaten als Parameter übergeben werden. Der Server analysiert dann mit Hilfe von räumlichen Datenbankabfragen (PostGIS) die Umgebung des Punktes und liefert ein strukturierteres JSON-Ergebnis, das Adresse, Ortsnamen und weitere Informationen enthält. Parameter wie zoom (für Detailgrad) oder addressdetails (für Adressstruktur) können verwendet werden, um das Ergebnisformat anzupassen. In dieser Arbeit wird Reverse Geocoding verwendet, um den durch GPS erfassten Standort der Kameraeinheit in eine Adresse umzuwandeln. Diese Adresse wird zusammen mit dem Parkstatus, dem Zeitstempel und den Koordinaten in einer Cloud-Datenbank gespeichert. So können Nutzer die Informationen nicht nur als

Koordinaten, sondern auch in verständlicher Form (z. B. „Bahnhofstraße 12, Nürtingen“) angezeigt bekommen. Durch diese Ergänzung wird die Funktionalität des Systems deutlich verbessert, da es nicht nur technisch präzise, sondern auch benutzerfreundlich arbeitet. [1]

Cloud-Anbindung

Um erkannte freie Parkplätze automatisiert und benutzerfreundlich an Endnutzer weiterzugeben, wird eine cloudbasierte Lösung integriert. Sobald das System mithilfe der Bildverarbeitung und GPS-Ortung einen freien Parkplatz erkennt, werden die zugehörigen Standortdaten – bestehend aus GPS-Koordinaten sowie der zugehörigen Adresse (Straße, Hausnummer und Stadt) – an eine Firebase Realtime Database übertragen. Firebase wurde als Cloud-Plattform gewählt, da sie eine einfache und effiziente Möglichkeit bietet, Daten in Echtzeit zwischen Geräten und Anwendungen zu synchronisieren. ## FicodeTechnologiesLimited[2](80) Über die integrierte REST API kann das System direkt vom Raspberry Pi aus Daten an die Cloud senden, ohne dass ein zusätzlicher Server notwendig ist. Dadurch wird die Systemarchitektur schlank gehalten und die Kommunikation effizient umgesetzt. Die übertragenen Daten beinhalten: ~~* ~ Breiten- und Längengrad (GPS) ~ Zugehörige Adresse (per Reverse Geocoding ermittelt) ~ Zeitstempel der Erkennung ~~~ Die mobile App kann anschließend diese Daten in Echtzeit abrufen und dem Endnutzer auf einer Karte oder in Listenform anzeigen. Durch die Nutzung der Firebase-Plattform sind spätere Erweiterungen wie Benachrichtigungen, Benutzerverwaltung oder Zugriffsbeschränkungen problemlos integrierbar. Diese Lösung ermöglicht eine zuverlässige, skalierbare und leicht wartbare Datenbereitstellung für smarte Parkplatzanwendungen mit direkter Anbindung zwischen Edge-Gerät (Raspberry Pi) und der Nutzeroberfläche.

Solarbetrieb für den Raspberry Pi Zero 2 WH

Für die autarke Energieversorgung meines Projekts mit dem Raspberry Pi Zero 2 WH habe ich das Sandberg Solar Ladegerät mit 21 W Solarpanel und integrierter 10.000 mAh Powerbank als ideale Lösung ausgewählt. Dieses Gerät vereint eine leistungsstarke Energieerzeugung durch hocheffiziente Solarpanels (23,5 % Wirkungsgrad) mit einer robusten, wetterfesten Powerbank, die sich für den Dauerbetrieb im Außenbereich eignet. Bei optimaler Sonneneinstrahlung liefert das faltbare 21-Watt-Panel ausreichend Energie, um den täglichen Strombedarf des Raspberry Pi zu decken – selbst bei einem energieintensiveren Betriebsprofil, bei dem das Gerät jede Minute für 10 Sekunden aktiv ist. Der integrierte Li-Polymer-Akku mit 10.000 mAh Kapazität dient als zuverlässiger Puffer bei wechselnden Lichtverhältnissen und ermöglicht dank Power-Through-Funktion den gleichzeitigen Betrieb und das Laden des Systems. Mit drei USB-Ausgängen (inkl. USB-C Power Delivery) bietet das System die nötige Flexibilität, um den Raspberry Pi effizient und stabil mit Energie zu versorgen. Die kompakte Bauweise sowie die Möglichkeit, das Panel unterwegs am Rucksack zu befestigen, machen dieses System besonders geeignet für mobile, netzunabhängige Anwendungen im Outdoor-Bereich. ## SandbergWorld[2](100)

Funktionsweise

Das Gesamtsystem der intelligenten Parkraumüberwachung basiert auf einem modularen Aufbau, der die Erfassung, Verarbeitung und Übertragung von Parkinformationen vollständig automatisiert abbildet. Herzstück ist ein Raspberry Pi Zero 2 WH, der über eine angeschlossene Kamera periodisch Bilder von einem Straßenabschnitt aufnimmt. Diese Bilder werden lokal mit einem Objekterkennungsmodell (z. B. YOLOv8) analysiert, um das Vorhandensein oder Fehlen von Fahrzeugen zu ermitteln. Erkennt das System, dass ein Fahrzeug über einen bestimmten Zeitraum (z. B. 15 Minuten) an einer Stelle steht, wird diese Position als belegt gespeichert. Sobald das Objekt verschwindet, wird der Parkplatz als frei gemeldet. Gleichzeitig erfasst ein angeschlossenes GPS-Modul die geografische Position des Systems. Die ermittelten Koordinaten werden durch Reverse Geocoding in eine lesbare Adresse umgewandelt. Alle relevanten Daten – bestehend aus Parkstatus, GPS-Position, Adresse und Zeitstempel – werden anschließend über eine Internetverbindung in eine Cloud-Datenbank hochgeladen. Dadurch können externe Systeme oder Benutzer in Echtzeit auf die Daten zugreifen. Die Energieversorgung erfolgt vollständig über ein mobiles Solarpanel mit integrierter Powerbank, das auch bei unsteten Lichtverhältnissen einen stabilen Betrieb ermöglicht. Alle Komponenten sind in einem wetterfesten Gehäuse untergebracht, um einen langfristigen Einsatz im Außenbereich zu gewährleisten.

Ausblick

Im Rahmen dieser Arbeit wurde ein mobiler, autark betriebener Prototyp zur intelligenten Parkraumüberwachung erfolgreich erweitert und umgesetzt. Die Integration von GPS, Cloud-Datenübertragung und solargestützter Energieversorgung macht das System flexibel einsetzbar – unabhängig von fest installierter Infrastruktur. In zukünftigen Weiterentwicklungen könnten zusätzliche Funktionen implementiert werden, etwa die Nutzung maschinellen Lernens zur Prognose der Parkplatzverfügbarkeit auf Basis historischer Daten

oder die Integration eines Energiemanagementsystems zur noch effizienteren Nutzung der Solarenergie. Auch die Einbindung einer Benutzeroberfläche in Form einer mobilen App oder Webplattform könnte realisiert werden, um Echtzeitinformationen über freie Parkplätze direkt an Endnutzer weiterzugeben. Zusätzlich wäre eine Erweiterung des Systems um weitere Sensorik (z. B. Umwelt- oder Bewegungssensoren) denkbar, um die Zuverlässigkeit zu erhöhen oder den Anwendungsbereich auszuweiten. Insgesamt zeigt dieses Projekt, dass eine intelligente, energieautarke und vernetzte Parkraumüberwachung mit vergleichsweise einfachen Mitteln technisch realisierbar ist und großes Potenzial für den städtischen Einsatz bietet.

Literatur und Abbildungen

[1] Danial Small. Reverse Geocoding. https://nominatim.org/release-docs/latest/api/Reverse/, 2009.

Automatisierte SBOM-Integration in CI/CD-Pipelines zur Echtzeit-Pflege von IT-Inventaren in LeanIX

Martin Derek **Dieter Morgenroth**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

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

Einleitung

Der bevorstehende Geltungsbeginn der EU-NIS-2-Richtlinie [4] verpflichtet Unternehmen, für jede ausgelieferte Software eine **Software Bill of Materials** (SBOM) vorzuhalten. Parallel müssen Enterprise-Architekten in SAP LeanIX ihre Microservice-Landschaft aktuell halten, um Lizenz- und CVE-Risiken sofort bewerten zu können. In vielen Projekten geschieht diese Pflege jedoch noch manuell in Excel-Listen oder Wikis. Diese sind bereits nach dem nächsten Commit veraltet.
Um diesen Medienbruch zu beseitigen, entwirft die Bachelorarbeit eine End-to-End-Pipeline, die vom Git-Commit bis zum aktualisierten LeanIX-Fact-Sheet alle Schritte automatisiert:

1. **Configuration-as-Code** – jedes Repository enthält ein Manifest ‘leanix.yaml‘, das Namen, Tags, Teams und Business-Applikationen des Microservices beschreibt.
2. **SBOM-Erzeugung** – ein GitLab-Runner generiert bei jedem Build mit Trivy ein SBOM im CycloneDX-Format [2].
3. **LeanIX-Ingestion** – ein Python-Script authentifiziert sich via OAuth 2.0 und überträgt Manifest und SBOM an die Self-Built-Software-API von LeanIX [3].
4. **Transparenzgewinn** – LeanIX reichert die Pakete automatisch mit CVE- und Lizenzinformationen an; Architekten sehen Risiken und Technologie-Standards in Echtzeit.

Diese Pipeline reduziert den manuellen Pflegeaufwand auf **null Stunden pro Release**, macht Open-Source-Risiken sofort sichtbar und schafft eine belastbare Datengrundlage für strategische Technologieentscheidungen.

Zielsetzung der Arbeit

Die Arbeit verfolgt das Ziel, einen vollständig automatisierten Workflow zu realisieren, der Microservices aus dem Quell-Repository heraus identifiziert, ihre Software-Stückliste erzeugt und den aktuellen Architektur-Stand in SAP LeanIX ohne manuelle Eingriffe synchron hält. Im Fokus steht der Web-Service „Probandenpool“ als Pilot.
Teilziele:

- **Manifest-Standardisierung** – Entwicklung eines YAML-Schemas (‘leanix.yaml‘) zur einheitlichen Beschreibung von Service-Metadaten, Tags, Teams und Business-Bezügen.
- **CI-basierte SBOM-Generierung** – Integration eines Trivy-Jobs in GitLab, der bei jedem Commit eine CycloneDX-SBOM erstellt und als Artefakt bereitstellt.
- **LeanIX-Ingestion-Script** – Implementierung eines Python-Tools, das OAuth 2.0 nutzt, Manifest und SBOM an die LeanIX-Self-Built-Software-API überträgt und Erfolgs-Logs ausgibt.
- **Sicherheits- und Lizenz-Transparenz** – Nachweis, dass LeanIX CVE- und Lizenzinformationen automatisiert anreichert und in Explorer-Dashboards darstellt.
- **Evaluierung der Effizienz** – Messung der Durchlaufzeit, des manuellen Pflegeaufwands und der SBOM-Abdeckung vor und nach Einführung der Pipeline.

Durch das Erreichen dieser Ziele wird gezeigt, dass eine Configuration-as-Code-Strategie nicht nur den Pflegeaufwand eliminiert, sondern auch die Sicherheits- und Entscheidungsqualität in Cloud-Architekturen maßgeblich erhöht.

Pipeline-Architektur

Die automatisierte Discovery-Kette besteht aus drei GitLab-Jobs:

- **generate-sbom** zieht ein Alpine-basiertes Trivy-Image aus und speichert 'sbom.json' als Artefakt.
- **leanix-discovery** lädt Manifest ('leanix.yaml') und SBOM via Python-Script hoch. Das Script nutzt den OAuth-Flow *apitoken* → *access_token*, ruft anschließend 'PUT /manifests' und 'POST /sboms' auf und loggt die FactSheet-ID.
- **Downstream-Jobs** Terraform, Docker, Destroy bleiben unverändert und werden über 'PIPELINE_TYPE' gesteuert.

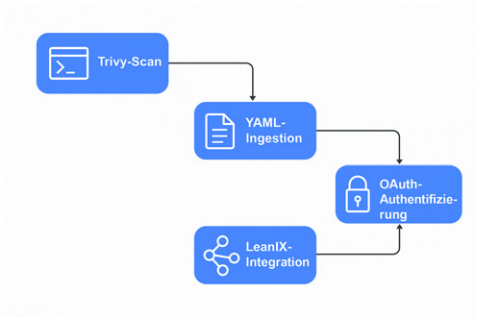
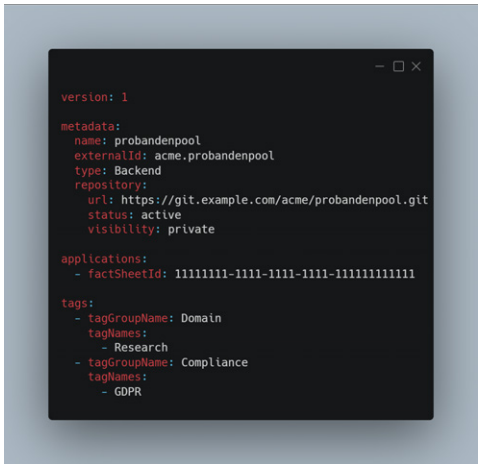


Abb. 1: Darstellung der CI/CD-Pipeline zur automatisierten SBOM-Erzeugung und LeanIX-Integration bestehend aus Trivy-Scan, YAML-Ingestion und OAuth-Authentifizierung. [1]

Die Jobs laufen in einem GitLab-Runner mit OIDC-Föderation zu AWS; Secrets wie 'LEANIX_API_TOKEN' sind als protected variables hinterlegt.

Implementierung

1 – Manifest Das YAML-Schema nutzt ‘metadata‘, ‘applications‘, ‘teams‘ und ‘tags‘. Simplifiziertes Beispiel:



```yaml
version: 1

metadata:
  name: probandenpool
  externalId: acme.probandenpool
  type: Backend
  repository:
    url: https://git.example.com/acme/probandenpool.git
    status: active
    visibility: private

applications:
  - factSheetId: 11111111-1111-1111-1111-111111111111

tags:
  - tagGroupName: Domain
    tagNames:
      - Research
  - tagGroupName: Compliance
    tagNames:
      - GDPR
```

Abb. 2: Codeausschnitt der Manifest Datei [1]

2 – SBOM-Job GitLab-Snippet:

```yaml
generate-sbom:
  image:
    name: aquasec/trivy:latest
    entrypoint: [""]
  script:
    - trivy fs --skip-db-update --format cyclonedx --output "$SBOM_FILE" .
  artifacts:
    paths: [ $SBOM_FILE ]
```

Abb. 3: Codeausschnitt der Pipeline [1]

3 – Python-Script (Ausschnitt)

```python
response = requests.post(
  LEANIX_OAUTH2_URL,
  auth=("apitoken", os.getenv("LEANIX_API_TOKEN")),
  data={"grant_type": "client_credentials"},
)
os.environ["LEANIX_ACCESS_TOKEN"] = response.json()["access_token"]
```

Abb. 4: Codeausschnitt des authentifizierungs Python Skript. [1]

4 – Workflow-Rules Diese stellen sicher, dass SBOM-/Discovery-Jobs auch auf Feature-Branches laufen, während Terraform-Jobs nur auf geschützten Branches ausgeführt werden.

Ergebnisse und Prognosen

- **Zeitersparnis** Pflegeaufwand für den Pilot-Service sank von Ø 1 h/Monat auf **0 h** (reine CI-Laufzeit 2 min).
- **Abdeckung** ‘sbom.json‘ listet ca. 125 Pakete (davon 78 Open-Source-Libraries). 100 % werden im LeanIX-Explorer angezeigt.

- **Security-Insights** LeanIX markiert 5 HIGH-CVEs sofort nach Upload und wurden in der Folge aktualisiert.
- **Datenqualität** Relations-Explorer zeigt automatisch IT-Products. Application-Link wurde korrekt aus Manifest übernommen.
- **Kosten-/Nutzen-Faktor** Einrichtungsaufwand einmalig ~2 Personentage; erwartete Einsparung >40 h/Jahr.

Ausblick

Zur weiteren Erhöhung der Sicherheitsautomatisierung ist vorgesehen, ein sogenanntes Vulnerability-Gate zu implementieren. Dieses würde die Pipeline automatisch abbrechen, sobald Trivy beim SBOM-Scan HIGH- oder CRITICAL-CVEs entdeckt (trivy sbom –exit-code 1). Dadurch könnten Sicherheitsrisiken bereits beim Build blockiert und frühzeitig adressiert werden.
Darüber hinaus ist ein Roll-out des Workflows auf alle Porsche-Cloud-Repositories geplant. Damit würde ein einheitlicher, CI-gestützter Discovery-Standard etabliert, der sowohl die Datenqualität als auch die Reaktionsgeschwindigkeit bei Lizenz- und Sicherheitsfragen unternehmensweit verbessert.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] OWASP Foundation. CycloneDX Specification Version 1.6. https://cyclonedx.org/specification/overview/, 03 2025.

[3] SAP SE. SAP LeanIX – Self-Built Software Discovery API (Version 2025-03). https://docs-eam.leanix.net/docs/self-built-software-discovery, 03 2025.

[4] Europäische Union. Richtlinie (EU) 2022/2555 über Maßnahmen für ein hohes gemeinsames Cybersicherheitsniveau. https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX:32022L2555, 12 2022.

Disaster Recovery - Soll/Ist Abgleich

Dersim Dogan **Thomas Rodach**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Bürgschaftsbank Baden-Württemberg GmbH, Stuttgart

Problemstellung

In Zeiten zunehmender Digitalisierung und Vernetzung ist die Gewährleistung der Geschäftskontinuität in Banken und Finanzinstituten von entscheidender Bedeutung. Dadurch stehen Finanzinstitute wie die Bürgschaftsbank Baden-Württemberg vor der Herausforderung, ihre kritischen Geschäftsprozesse und IT-Systeme vor Ausfällen effektiv zu schützen, um so einen reibungslosen Geschäftsbetrieb weiterhin sicherstellen zu können. IT-Ausfälle und unzureichende Notfallmaßnahmen können dabei zu erheblichen finanziellen Schäden, langfristigen Reputationsverlusten und weitreichenden regulatorischen Konsequenzen führen. Daher ist der Einsatz eines umfassenden und wirksamen DR Konzepts unerlässlich. Im Krisenfall soll eine schnelle, sichere und zuverlässige Wiederaufnahme der kritischen Geschäftsprozesse gewährleistet werden müssen. Hierzu existieren zwar theoretisch fundierte und standardisierte Vorgehensweisen, die in Best-Practices und regulatorischen Richtlinien beschrieben werden. Diese finden jedoch in der praktischen Umsetzung oftmals einen deutlichen Unterschied, zwischen den formellen Vorgaben und der tatsächlichen Umsetzung [3]. Vor diesem Hintergrund ist es von hoher Relevanz, dass die Bürgschaftsbank Baden-Württemberg ein detailliertes und differenziertes Verständnis über den aktuellen Umsetzungsgrad ihres Disaster-Recovery-Konzepts gewinnt. Die systematische Identifikation und tiefgehende Analyse bestehender Schwachstellen sowie die Bewertung potenzieller Risiken sind essenziell, um den Anforderungen an die Notfallplanung und das Notfallmanagement optimal gerecht zu werden. Ein klar strukturierter Soll/Ist-Abgleich bietet dabei eine Grundlage, um etwaige Mängel zu ausfindig zu machen und Handlungsempfehlungen konkret aussprechen. Die Ergebnisse einer solchen systematischen Analyse können entscheidend dazu beitragen, bestehende Prozesse kritisch zu hinterfragen, gezielt Optimierungen vorzunehmen und effektive präventive Maßnahmen zu implementieren, um somit langfristig die operative Widerstandsfähigkeit und Resilienz der Bürgschaftsbank nachhaltig zu stärken [2].

Zielsetzung der Arbeit

Ziel der vorliegenden Arbeit ist die Analyse des bestehenden DR Konzepts der Bürgschaftsbank Baden-Württemberg im Hinblick auf seine Wirksamkeit, organisatorische Umsetzung und regulatorische Konformität. Im Zentrum steht ein strukturierter Soll-/Ist-Abgleich, der klären soll, inwieweit interne Richtlinien, aufsichtsrechtliche Anforderungen (insbesondere gemäß BAIT und MaRisk) sowie bewährte Best Practices im praktischen Betrieb umgesetzt sind. Der Fokus der Untersuchung liegt dabei auf den organisatorischen und prozessualen Komponenten des Notfallmanagements – insbesondere auf Rollenverteilungen, Dokumentationen, Testverfahren und Kommunikationsstrukturen. Zur Beurteilung der tatsächlichen Handlungsfähigkeit im Ernstfall werden zusätzlich strukturierte Interviews mit verantwortlichen Fachbereichen geführt. Ziel ist es, bestehende Schwachstellen in der Notfallorganisation zu identifizieren und auf dieser Basis fundierte Handlungsempfehlungen zur Optimierung der Prozesse abzuleiten. Die Arbeit soll damit einen praxisnahen Beitrag zur Stärkung der organisatorischen Resilienz und zur Weiterentwicklung des Notfallmanagements leisten.

Theoretische Grundlagen

Bevor konkrete Zielgrößen wie RTO und RPO betrachtet werden können, ist es notwendig, zwischen unterschiedlichen Eskalationsstufen eines IT-Vorfalls zu unterscheiden. Abbildung 1 zeigt die Einordnung eines Ereignisses als Störung, Notfall oder Krise im Kontext des Business Continuity Managements.

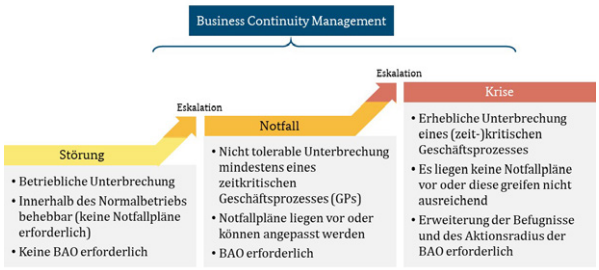


Abb. 1: Abgrenzung zu Störung, Notfall & Krise [1]

Ein effektives DR Konzept basiert auf klar definierten Zielgrößen, die im Notfall eine strukturierte Wiederherstellung kritischer IT-Services ermöglichen. Die wichtigsten Kennzahlen in diesem Kontext sind:

- **Recovery Point Objective (RPO):** Gibt den maximal tolerierbaren Datenverlust an – also die Zeitspanne zwischen dem letzten gesicherten Datenstand und dem Schadensereignis.
- **Recovery Time Objective (RTO):** Die Zeitspanne, innerhalb der ein System nach einem Ausfall wieder betriebsfähig sein muss, um signifikante Schäden zu vermeiden.

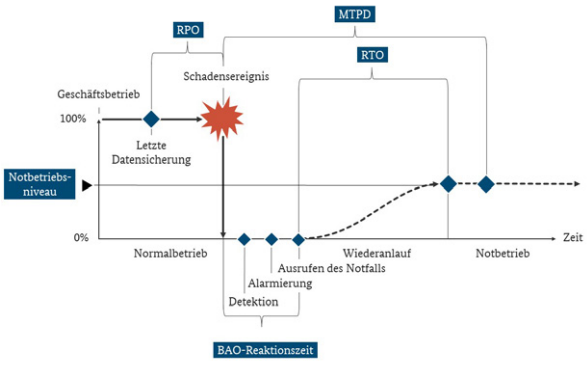


Abb. 2: Erläuterung der Kennzahlen [1]

- **Maximum Tolerable Period of Disruption (MTPD):** Der längste Zeitraum, den ein Geschäftsprozess unterbrochen sein darf, ohne dass die Organisation ernsthaft gefährdet wird.
- **Notbetriebsniveau:** Das minimale Betriebsniveau, das während der Wiederanlaufphase gewährleistet sein muss.
- **BAO-Reaktionszeit:** Zeitraum zwischen Detektion des Vorfalls und dem organisierten Übergang in den Notbetrieb, einschließlich Alarmierung und Ausrufen des Notfalls.

Diese Parameter sind essenziell für die Planung und Bewertung von Wiederanlaufstrategien. Abbildung 2 zeigt den zeitlichen Ablauf eines Notfallszenarios und ordnet die genannten Begriffe grafisch ein.
Ein Verständnis dieser Größen bildet die Grundlage für die Analyse bestehender DR-Konzepte sowie für die Bewertung der organisatorischen Handlungsfähigkeit im Ernstfall.

Fazit

Die Analyse des bestehenden DR Konzepts der Bürgschaftsbank Baden-Württemberg zeigt, dass organisatorische Maßnahmen und einheitliche Abläufe essenziell für ein wirksames Notfallmanagement sind. Insbesondere der strukturierte Soll-/Ist-Abgleich hat verdeutlicht, dass zwischen formell vorhandenen Konzepten und deren praktischer Umsetzung teils deutliche Abweichungen bestehen. Um die Handlungsfähigkeit im Ernstfall sicherzustellen, sind regelmäßige Tests, eine klare Definition von Rollen und Verantwortlichkeiten sowie ein gemeinsames Verständnis über kritische Abläufe von zentraler Bedeutung. Ergänzend sollte die Notfalldokumentation aktuell gehalten und zentral zugänglich sein. Die gewonnenen Erkenntnisse liefern eine fundierte Grundlage, um bestehende Prozesse kritisch zu hinterfragen und gezielte Verbesserungsmaßnahmen abzuleiten, die zur Stärkung der organisatorischen Resilienz beitragen.

Literatur und Abbildungen

[1] Bundesamt für Sicherheit in der Informationstechnik. *Business Continuity Management - BSI-Standard 200-4*, volume 1. Bundesamt für Sicherheit in der Informationstechnik, 1 edition, 2023.

[2] Christian Nern, Julian Krautwald, and Robert Ullrich. Globale IT-Ausfälle: Ein aktueller Weckruf für Finanzunternehmen. https://klardenker.kpmg.de/financialservices-hub/globale-it-ausfaelle-ein-aktueller-weckruf-fuer-finanzunternehmen/, 07 2024.

[3] Enterprise Strategy Group Veeam Software. Veeam Availability Report: Why organizations still struggle to digitally transform and what they can do about it. https://branden.biz/wp-content/uploads/2018/03/2017_availability_report-1.pdf, 03 2017.

Robustheit von LMDrive bei variablen Umweltbedingungen

Luca Effenberger **Markus Enzweiler**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Die Entwicklung autonomer Fahrzeuge hat in den letzten Jahren beachtliche Fortschritte gemacht. Viele Systeme setzen dabei auf eine modulare Architektur, bei der Aufgaben wie Wahrnehmung, Entscheidungsfindung, Trajektorienvorhersage und Mensch-Fahrzeug-Interaktion in spezifische Komponenten unterteilt werden. Diese strukturierte Herangehensweise erlaubt eine spezialisierte Entwicklung und gezielte Optimierung einzelner Module [3]. Im Gegensatz dazu verfolgen moderne End-to-End-Systeme einen integrierten Ansatz: Ein einzelnes Modell verarbeitet die Sensordaten und generiert direkt Steuerentscheidungen. Diese holistische Struktur soll Fehlerfortpflanzung reduzieren und die Reaktion auf dynamische Umgebungen verbessern. Durch das Lernen aus Daten kann das System gesamtheitlich optimiert werden, anstatt isolierte Komponenten zu verbessern [3]. Besonders innovativ ist dabei das Projekt LMDrive [2], das große Sprachmodelle (LLMs) für die direkte Fahrzeugsteuerung in einer geschlossenen Rückkopplungsschleife (Closed-Loop) einsetzt 1. In dieser Evaluationsform fließen die vom Modell generierten Steuerbefehle direkt in die Simulation ein, wodurch seine Entscheidungen unter realitätsnahen Bedingungen getestet werden können. Kern des LMDrive-Systems ist die Kombination aus visueller Wahrnehmung und sprachbasierter Entscheidungslogik. Multimodale Sensordaten – insbesondere Kamerabilder und LiDAR-Signale – werden durch einen Vision Encoder in sogenannte „visuelle Tokens“ übersetzt, die eine abstrahierte Darstellung der Umgebung liefern. Diese Tokens werden gemeinsam mit einer Spracheingabe – z. B. „Biegen Sie an der nächsten Kreuzung links ab“ – an das Sprachmodell übergeben. Das LLM verarbeitet diese Informationen sequentiell und generiert in Textform eine Fahrzeuganweisung, die anschließend in konkrete Steuerbefehle wie Lenk- oder Bremssignale übersetzt wird.

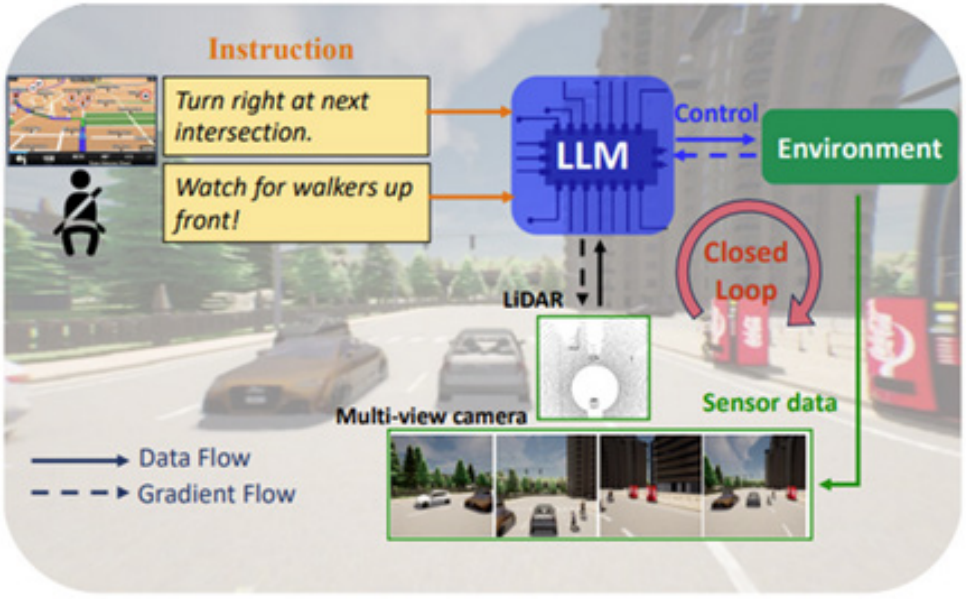


Abb. 1: Der Aufbau von LMDrive [2]

Evaluation im CARLA-Simulator

Getestet wird LMDrive mithilfe des Open-Source-Simulators CARLA [1], der eine realitätsnahe städtische Umgebung mit verschiedenen Fahrzeugen, Verkehrsteilnehmern, Wetterlagen und konfigurierbaren Sensorsystemen wie Kamera, LiDAR und Radar bietet. Dank seiner hohen Flexibilität eignet sich CARLA besonders gut zur Entwicklung, Validierung und Evaluation autonomer Fahralgorithmen in den Bereichen Wahrnehmung, Planung und Steuerung [3]. Ein zentraler Aspekt der Evaluation besteht darin zu untersuchen, wie robust LMDrive gegenüber unterschiedlichen Umweltbedingungen ist. Für menschliche Fahrer stellen Faktoren wie Regen, Nebel oder Dunkelheit eine erhebliche Herausforderung dar, da sie sowohl die Sichtverhältnisse als auch das Fahrverhalten beeinflussen. Ziel der Untersuchung ist es daher, herauszufinden, wie sich solche Bedingungen auf die Leistungsfähigkeit des LMDrive-Modells auswirken.

Abb. 2: Unterschiedliche Umweltbedingungen im CARLA-Simulator [1]

Hierzu werden Benchmarks erstellt, die aus vordefinierten Routen unter verschiedenen Wetterbedingungen bestehen 2. CARLA bietet dafür diverse urbane Szenarien und die Möglichkeit, über Wegpunkte reproduzierbare Fahrstrecken zu definieren. Jede Route wird mehrfach bei unterschiedlichen Wetterlagen durchfahren, wobei CARLA standardisierte Wetterszenarien zur Verfügung stellt, darunter Parameter wie Regenintensität, Nebeldichte und Tageszeit. Während der Fahrten erzeugt CARLA eine Ausgabedatei, die unter anderem dokumentiert, ob die Route erfolgreich absolviert wurde und ob es zu Kollisionen kam. Auf Basis dieser Daten lässt sich analysieren, inwieweit Umweltbedingungen die Funktionsweise von LMDrive beeinflussen.

Ausblick

Die in dieser Arbeit gewonnenen Erkenntnisse bieten zudem eine fundierte Grundlage für gezielte Verbesserungen von LMDrive, sofern diese notwendig sein sollten. Ein nächster Schritt könnte darin bestehen, das System durch weiteres Finetuning robuster gegenüber schwierigen Umweltbedingungen zu machen – etwa bei Regen oder schlechter Sicht. Denkbar wäre dabei ein Training der vorgeschalteten Komponenten wie der Q-Former und des Visual Encoders [2]. Die dafür benötigten Trainingsdaten könnten direkt aus CARLA stammen, indem man Szenen auswählt, in denen das Modell Schwächen zeigt. Ein vergleichbarer Ansatz wird auch im Projekt DriveGPT4 angestrebt, das ebenfalls CARLA-basierte Daten zur Verbesserung plant [4].

Literatur und Abbildungen

[1] A Dosovitskiy, G Ros, F Codevilla, A Lopez, and V Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning (CoRL 2017)*. Sergey Levine, Vincent Vanhoucke, Ken Goldberg, 2017.

[2] H Shao, Y Hu, L Wang, G Song, S L Waslander, Y Liu, and H Li. Lmdrive: Closed-loop end-to-end driving with large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024)*. IEEE, 2024.

[3] H Tian, K. Reddy, Y Feng, M Quddus, Y Demiris, and P Angeloudis. Large (Vision) Language Models for Autonomous Vehicles: Current Trends and Future Directions. *Authorea Preprints*, 2024.

[4] Z Xu, Y Zhang, E Xie, Z Zhao, Y Guo, K Y K Wong, Z Li, and H Zhao. DriveGPT4: Interpretable End-to-End Autonomous Driving Via Large Language Model. *IEEE Robotics and Automation Letters*, 2024.

Kryptografische Operationen und sichere Ausführung auf i.MX8M Nano

Halit Osman Efkere **Markus Enzweiler**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Festool GmbH, Wendlingen

Zielsetzung

Heutzutage sind Elektrogeräte wie Kühlschränke, Zahnbürsten, Kaffeeautomaten, Smartwatches, Smartphones und Elektrowerkzeuge zunehmend vernetzt und werden dadurch zu IoT-Geräten. Sie verfügen über Schnittstellen wie WLAN oder Bluetooth, um Dienste wie Gerätesteuerung und Statistiken für den Endbenutzer bereitzustellen sowie Software-Updates zu erhalten. Es werden sensible Daten gespeichert und verarbeitet, darunter WLAN-Passwörter, Sicherheitszertifikate und persönliche Daten. Infolgedessen sind IoT-Geräte ein Ziel für Cyberangriffe. Zugleich sind die Ressourcen in einem eingebetteten System begrenzt; die Recheneinheit in einem Kühlschrank mit einem kleinen Display und einem Mikrocontroller mit WLAN-Schnittstelle kann deshalb nur einfache Aufgaben übernehmen, was die Umsetzung von Sicherheitsmaßnahmen erschwert. Dabei muss die Sicherheit der Geräte bei begrenzter Rechenleistung besonders beachtet werden [4].

Die zunehmende Anzahl von IoT-Geräten führt zu neuen Anforderungen an die Cybersicherheit, da diese Geräte im Fall von Cyberangriffen für den Verbraucher kritische Auswirkungen haben. Am 23. Oktober 2024 wurde daher vom Europäischen Parlament und Rat die Verordnung zum Cyber-Resilience-Act (CRA) verabschiedet, die für Produkte mit digitalen Elementen (Software) Sicherheitsmaßnahmen und Anforderungen vorschreibt. Beispielsweise sollen die Integrität und Vertraulichkeit der Daten durch Verschlüsselung geschützt werden [5]. Ziel der Arbeit ist aufzuzeigen, wie die Anforderungen auf der Hardware- und Softwareebene unter Berücksichtigung der Hardwareplattform NXP i.MX8M Nano umgesetzt werden können. Dieser Artikel gibt einen Überblick über die dazu verwendeten Subsysteme.

Der System-on-Chip i.MX8M Nano

Der Gegenstand dieser Untersuchung ist der System-on-Chip (SoC) von NXP mit zwei Cortex-A-Kernen auf Basis der ARM-Architektur. Der SoC bietet Unterstützung für 3D-Beschleunigung (OpenGL), Ethernet-Schnittstelle, I²C, SPI und UART, wie in Abbildung 1 dargestellt, sowie Sicherheitskomponenten, welche das Hauptthema dieser Arbeit sind. Die Sicherheitsarchitektur ARM-TrustZone wird vom SoC für eine sichere Ausführungsumgebung auf der Hardwareebene unterstützt – ebenso ein sicherer Speicherbereich (Secure Non-Volatile Storage), der eine gerätespezifische Identifikationsnummer sowie einen Master-Key bereitstellt. Zudem stellt der SoC ein Hardwaremodul (CAAM) für Hardware-beschleunigte Verschlüsselungsaufgaben [6].

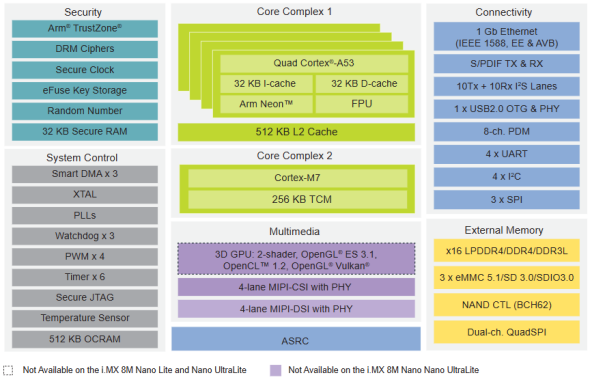


Abb. 1: Blockdiagramm i.MX8M Nano [6]

Cryptographic Accelerator and Assurance Module (CAAM)

Im SoC wird für die Ausführung der kryptografischen Operationen in der Hardware das Modul CAAM zur Verfügung gestellt. Abbildung 2 zeigt die Modularchitektur und Komponenten für kryptografische Operationen wie AES-256, RSA-4096 und Random-Number-Generator. Das Modul bietet Schnittstellen für das Betriebssystem, welche über drei Job-Ringe in der Hardware unabhängig voneinander ausgeführt werden können. Das CAAM hat eine direkte Verbindung zum

Secure-Non-Volatile Storage (SNVS), der für sicheres Speichern einen Master-Key liefert [7].

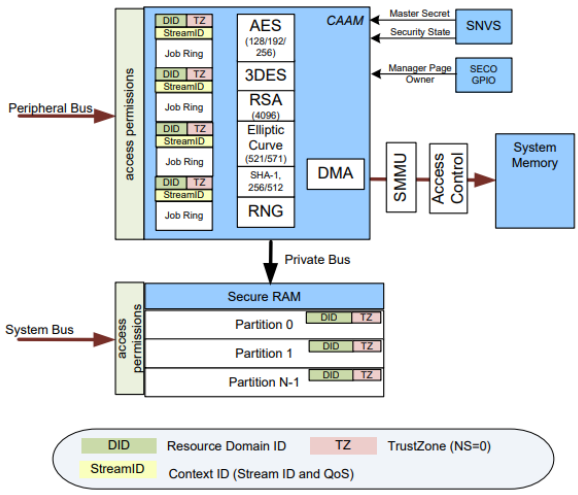


Abb. 2: CAAM-Modularchitektur [2]

Secure Non-Volatile Storage (SNVS)

Der sichere nichtflüchtige Speicher (SNVS) ist ein logischer Block im SoC, zuständig für die Überwachung und Speicherung sicherheitskritischer Daten (Master-Key). Dieser Master-Key wird beim Herstellungsprozess im SNVS als Fuse gebrannt. Er ist einzigartig für den jeweiligen SoC. Damit er verwendet werden kann, muss der Status sicher sein und es dürfen keine Sicherheitsverletzungen vorliegen. Dieser logische Block überwacht sicherheitsrelevante Ereignisse über den System-Security-Monitor, wie Tamper-Detection im Fall eines Außenzugriffs auf das Gerät und den SoC über die Tamper-PINs. In Abbildung 3 wird die interne Architektur des Blocks SNVS dargestellt. Dabei sind die Eingänge zum System-Security-Monitor sichtbar, die aus unterschiedlichen Quellen – etwa CAAM, Watchdogs und der Stromversorgung (gegen sogenannte Glitches) stammen. Mit diesen Eingängen definiert SNVS, ob sich der Chip im sicheren oder in einem unsicheren Status befindet, der auf Sicherheitsverletzungen zurückzuführen ist [7].

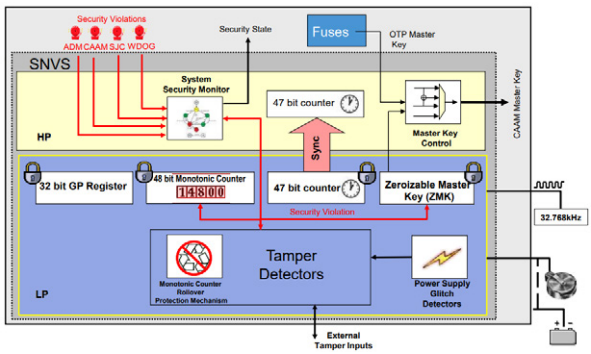


Abb. 3: SNVS-Sicherheitsarchitektur [2]

Trusted Execution Environment (TEE)

Neben den Sicherheitsmechanismen auf der Hardwareebene müssen auch Softwarekomponenten in der Lage sein, die Vertraulichkeit und Integrität der Daten zu gewährleisten. An dieser Stelle bietet die ARM-TrustZone-Sicherheitsarchitektur eine Isolation zwischen der sicheren Ausführungsumgebung (Secure-World) und der unsicheren Ausführungsumgebung (Normal-World). Die Isolation erfolgt auf der Hardwareebene durch Speicher- und IO-Schutz. Der Prozessor unterscheidet diese zwei Welten und aktiviert je nach Welt den Zugang zur Real-Time-Clock und zum Security-Controller oder deaktiviert diesen.
Als TEE wird die Open-Source-Portable-TEE (OP-TEE) im SoC neben dem Betriebssystem Linux eingesetzt. Ein wesentlicher Unterschied besteht darin, dass OP-TEE vor dem Linux-Kernel startet und in einem separaten Speicherbereich mit einer auf sicherheitskritische Aufgaben beschränkten Codebasis läuft. Somit laufen in OP-TEE die Trusted-Applikationen zur Verarbeitung von Signatur, Verschlüsselung, Entschlüsselung und Schlüsselgenerierung. Wenn das Hauptbetriebssystem – das unter anderem für die Bereitstellung der Benutzeroberfläche zuständig ist – eine Schwachstelle hat, ist OP-TEE davon nicht betroffen und bleibt isoliert.
Abbildung 4 schildert die Kommunikation zwischen Normal-World und Secure-World. Wenn eine Client-Applikation im Linux-User-Space einen Schlüssel von der Trusted-Applikation anfordert, wird die TEE-Client-API angesprochen und ein Schlüssel wird in der Trusted-Applikation generiert. Die Daten werden über Shared-Memory ausgetauscht [1].

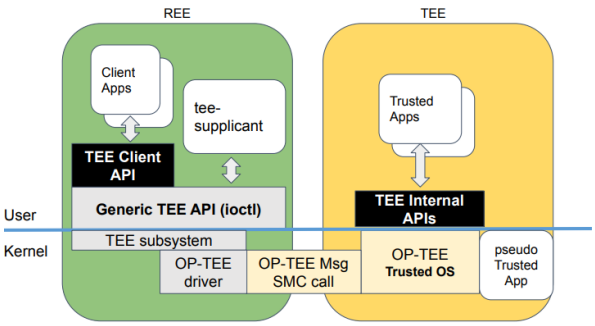


Abb. 4: Secure-World und Normal-World [3]

Ausblick

Die Sicherheitskomponenten des SoC i.MX8M Nano – CAAM, SNVS und OP-TEE – bieten eine solide Basis gegen Cyberangriffe. Die kryptografischen Operationen werden im Hardwaremodul CAAM ausgeführt und Aufgaben wie Schlüsselgenerierung werden in OP-TEE isoliert bearbeitet. Die Kombination aus CAAM, SNVS, ARM-TrustZone und OP-TEE ermöglicht die Umsetzung der Anforderungen des Cyber-Resilience-Acts. Infolgedessen werden die Vertraulichkeit und Integrität der Daten sowie die isolierte Ausführungsumgebung untersucht und umgesetzt.

Literatur und Abbildungen

[1] A. Bhat. Trusted Software Development Using OP-TEE. https://www.timesys.com/security/trusted-software-development-op-tee/, 2017.

[2] J. Cotner. i.MX 8 Security Overview. https://community.nxp.com/pwmxy87654/attachments/pwmxy87654/tech-days/271/1/AMF-AUT-T3363.pdf, 2018.

[3] S. Garg. Kernel TEE subsystem evolution. https://lpc.events/event/16/contributions/1295/attachments/981/1907/LPC22_%20Kernel%20TEE%20subsystem%20evolution.pdf, 2025.

[4] H. Ju et al. A Study on the Hardware-Based Security Solutions for Smart Devices. In *Proceedings of the 2015 International Conference on Computational Science and Computational Intelligence (CSCI)*. Institute of Electrical and Electronics Engineers, 2015.

[5] European Parliament and Council of the European Union. Verordnung (EU) 2024/2847 über Cybersicherheitsanforderungen für Produkte mit digitalen Elementen. http://data.europa.eu/eli/reg/2024/2847/oj/deu, 2024.

[6] NXP Semiconductors NV. i.MX8M Nano. https://www.nxp.com/products/i.MX8MNANO, 2019.

[7] NXP Semiconductors NV. *i.MX 8M Nano Applications Processor Reference Manual*. NXP Semiconductors NV, 2022.

Bildbasierte Knickwinkelerkennung für einen Sattelzug

Ben Engelhardt **Thomas Rothermel**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

Im Kontext des Fahrens von Lkw-Gespannen ist die präzise Erfassung des Knickwinkels, des relativen Winkels zwischen Zugmaschine und Trailer, von zentraler Bedeutung. Insbesondere bei Fahrmanövern wie dem Rückwärtsfahren oder dem Durchfahren enger Kurven. Herkömmliche Systeme zur Winkelmessung basieren zumeist auf physikalischen Sensoren, die sowohl an der Zugmaschine als auch am Anhänger installiert werden müssen. Dies führt zu erhöhtem technischem Aufwand, hohen Kosten und einer eingeschränkten Übertragbarkeit auf unterschiedliche Fahrzeugtypen. Diese Arbeit untersucht daher einen alternativen, bildbasierten Ansatz zur Knickwinkelschätzung mittels rückwärts gerichteter Kameras an der Fahrzeugseite, wie in Abbildung 1 dargestellt. Unter Verwendung eines digitalen und physischen Modellversuchs werden Bilddaten unter variierenden Konditionen aufgenommen, relevante Bildmerkmale extrahiert und zur geometrischen Winkelbestimmung herangezogen. Ziel ist die Entwicklung eines robusten, möglichst kalibrierungsarmen Verfahrens, das eine kostengünstige und flexible Alternative zu bestehenden hardwarebasierten Assistenzsystemen darstellen kann. Ein vergleichbarer videobasierter Ansatz wurde bereits für Pkw-Anhängerkombinationen erfolgreich implementiert und erreichte eine Schätzgenauigkeit von unter zwei Grad [1].

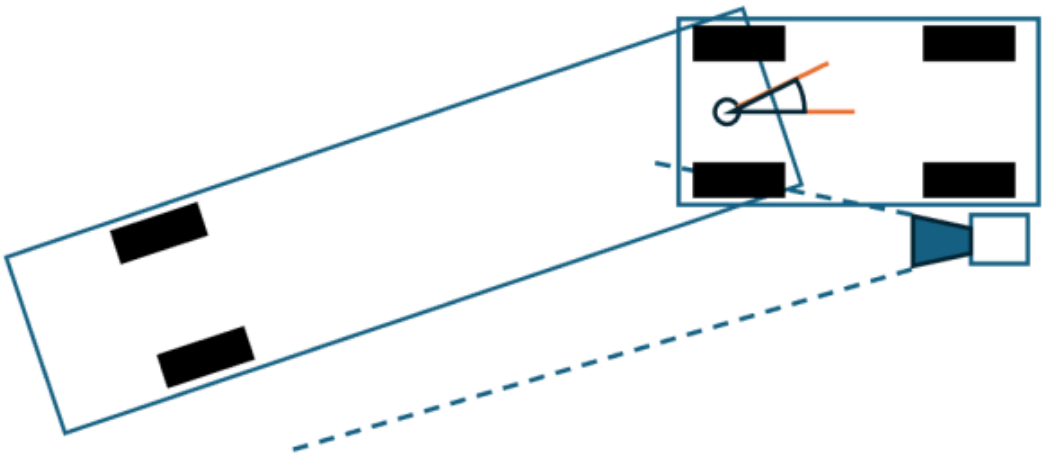


Abb. 1: Skizze des Versuchsaufbau, welche den Knickwinkel zwischen Zugmaschine und Trailer, sowie die angebrachten Kameras veranschaulicht. [3]

Problemstellung

Die bildbasierte Schätzung des Knickwinkels ist mit einer Vielzahl technischer Herausforderungen verbunden. So muss die Erkennung relevanter Bildmerkmale auch unter variierenden Lichtverhältnissen und unterschiedlichen Umgebungsbedingungen zuverlässig erfolgen. Darüber hinaus bedingt die Vielfalt an Fahrzeugtypen und Fahrzeug Geometrien den Einsatz adaptiver und generalisierbarer Algorithmen.

Arbeitsbeschreibung

Zur Untersuchung der Machbarkeit wird ein modularer Versuchsaufbau entwickelt, der sowohl physische als auch digitale Modellumgebungen berücksichtigt:

- **Digitales Modell (Simulation)**: In einer Simulationsumgebung wie CARLA können Sattelzüge modelliert und Kamerapositionen variabel festgelegt werden. Dies erlaubt die Generierung synthetischer Bilddatensätze unter kontrollierten Bedingungen (z. B. Licht, Wetter, Hintergrund). Die Simulation eignet sich besonders für die frühe Phase der Feature-Extraktion und zur systematischen Analyse unterschiedlicher Kameraperspektiven.
- **Physisches Modell**: Zur praxisnahen Validierung wird ein maßstabsgetreues Lkw-Modell aus Holz oder Kunststoff eingesetzt, bei dem eine angebrachte Kamera als bildgebendes System dient. Reale Umgebungsbedingungen und tatsächliche Knickwinkelmessungen bieten eine hohe Anwendungsnähe und ermöglichen die Validierung unter realistischen Störungen.

In beiden Fällen werden Bilddaten bei definierten Knickwinkeln aufgenommen und anschließend mit Python verarbeitet. Zur Feature-Extraktion dienen Kanten-, Linien- und Eckendetektion sowie ggf. visuelle Marker zur Vereinfachung der Erkennung. Alternativ wird auch der Einsatz von Deep-Learning-Methoden diskutiert, die jedoch einen umfangreichen, annotierten Datensatz erfordern, wie etwa im Ansatz von Dahal et al. [2]. Die geschätzten Bildfeatures werden durch geometrisches Mapping in eine Winkelabschätzung überführt. Ergänzend werden Machine-Learning-Verfahren in Betracht gezogen, die auf Basis der Bildmerkmale eine Regressionsfunktion zur direkten Winkelbestimmung lernen. Die Resultate werden abschließend mit den Referenzwerten verglichen und hinsichtlich Genauigkeit, Robustheit und Kameraposition bewertet.

Zielsetzung der Arbeit

Ziel der Arbeit ist es, ein funktionales Gesamtsystem zur bildbasierten Schätzung des Knickwinkels zwischen Zugmaschine und Trailer zu entwerfen und zu evaluieren. Dabei soll insbesondere untersucht werden:

- Welche Bildmerkmale sich zur zuverlässigen Winkelabschätzung eignen
- Wie verschiedene Kamerapositionen die Schätzgenauigkeit beeinflussen
- Inwieweit ein Verfahren ohne fahrzeugspezifische Kalibrierung auskommt
- Wie gut digitale Simulationsdaten auf reale Umgebungen übertragbar sind

Durch die Kombination von klassischer Bildverarbeitung und optionalem Machine Learning wird eine methodische Vergleichbarkeit sichergestellt.

Ausblick

Langfristig könnten bildbasierte Verfahren die Basis für kostengünstige und flexibel einsetzbare Zustandsdetektion bei Sattelzügen bilden – sei es zur Unterstützung des Fahrers bei kritischen Manövern oder als integraler Bestandteil autonomer Fahrfunktionen. Insbesondere durch den Einsatz synthetischer Bilddaten aus Simulationen lässt sich das Verfahren skalieren und auf neue Fahrzeugkonfigurationen übertragen. Weiterführende Arbeiten könnten die Robustheit gegenüber widrigen Bedingungen (Regen, Nacht, Verschmutzung der Kamera) erhöhen oder durch neuronale Netze Schätzung des Winkels ermöglichen.

Literatur und Abbildungen

[1] Lukas Caup et al. Video-Based Trailer Detection and Articulation Estimation. *Engineers, Institute of Electrical and Electronics Intelligent Vehicles Symposium, Proceedings*, 2013.

[2] Ashok Dahal et al. DeepTrailerAssist: Deep Learning Based Trailer Detection, Tracking and Articulation Angle Estimation on Automotive Rear-View Camera. *Engineers, Institute of Electrical and Electronics/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019.

[3] Eigene Darstellung.

Entwicklung einer API-Erweiterung zur clientseitigen UI-Generierung auf Basis von JSON Forms und HATEOAS

Nico Epp **Mirko Sonntag**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Adapt2Move GmbH, Augustinerstraße 22, 73728 Esslingen am Neckar

Einleitung

In den letzten Jahren hat sich erwiesen, dass flexible und skalierbare Architekturen entscheidend sind, um die sich ständig wechselnden Anforderungen zu erfüllen. Unter anderem dienen Web Application Programming Interfaces, oder besser bekannt als Web APIs zur Vernetzung und dem Austausch von Daten und als Grundlage der meisten Web Anwendungen, die heutzutage gang und gäbe in der modernen Softwareentwicklung sind. Weitgehend haben sich RESTful APIs zum Standard entwickelt, da sie eine unkompliziertere Anbindung erlauben und, auch im Vergleich zu anderen Technologien wie beispielsweise SOAP, ressourcen-freundlicher und auch umgänglicher im Datenhandling sind. [4]

Während APIs für das Bereitstellen von Daten und Vorgängen zuständig sind, liegt die Logik zur Erstellung der Benutzeroberfläche (UI) in der Regel auf der Client-Seite. Denn zum Darstellen der UI-Elemente mit Daten aus der API muss der Aufbau über diese Daten aus der API bekannt sein, um mit diesen arbeiten zu können. Das führt dazu, dass die UI an die API begrenzt ist und somit nicht wirklich komplett dynamisch agiert. Das Ziel dieser Arbeit ist es, eine API-Erweiterung zu gestalten, welche auf dem HATEOAS Ansatz basiert und die Benutzeroberflächen auf der Client-Seite mithilfe der Funktionalität von *JSON Forms* generieren soll.

Die Motivation für die Entwicklung einer API-Erweiterung, die HATEOAS mit clientseitiger UI-Generierung verknüpft, liegt primär in der Überwindung von Herausforderungen moderner Softwareentwicklung. Denn die klassischen, monolithischen Architekturen und geschlossenen Technologiestacks führen oftmals zu hohen Anpassungskosten und schränken die Wiederverwendbarkeit in dem Sinne der flexiblen und skalierbaren Anwendungsentwicklung ein.

So soll eine modulare und plattformunabhängige Architektur realisiert werden, die es so momentan noch nicht gibt.

Grundlagen

Representational State Transfer oder auch besser bekannt als REST ist ein Architekturstil, welcher auf dem Prinzip von Statelessness basiert, sowie der Verwendung von Standardmethoden auf Ressourcen. Im Vergleich zu anderen Protokollen wie SOAP, welche komplexe XML-basierte Nachrichtenformate und WSDL zur Beschreibung nutzen, wird bei RESTful Anwendungen besonders auf die Standardisierten HTTP-Methoden wie *GET*, *POST*, *PUT* und *DELETE* geachtet. Außerdem behandeln RESTful APIs Informationen als Ressourcen, die nur über eindeutige URIs zur Verfügung stehen und auch nur so zu erreichen sind. RESTful APIs sind einfach mit gängigen Programmen wie Postman oder über Methoden in JavaScript zugänglich und geben ihre Antwortdaten in gängigen Formaten wie JSON oder XML aus. [4]

HATEOAS (Hypermedia as the Engine of Application State) ist ein grundlegender REST-Architekturstil und stellt dem Client die vom Server gelieferten Hypermedia-Links (Hyperlinks) in Form eines weiteren JSON Objekts zur Verfügung, um z. B. durch die Anwendung zu navigieren, ohne oder nur mit sehr geringem Vorwissen darüber, wie genau die API aufgebaut ist. In der nachfolgenden Abbildung 1 ist eine beispielhafte Antwort im HATEOAS Schema dargestellt, in diesem man die auszugebenden weiterführenden Hyperlinks anschaulicher sieht. [3]

```
{
  "userId": "12345",
  "name": "John Doe",
  "email": "john.doe@acme.com",
  "links": [
    {
      "rel": "self",
      "href": "/users/12345"
    },
    {
      "rel": "edit",
      "href": "/users/12345/edit",
      "description": "Update the users details in the Database"
    },
    {
      "rel": "delete",
      "href": "/users/12345/delete",
      "description": "Remove the user from the Database"
    }
  ]
}
```

Abb. 1: Ausgabe von Daten mit HATEOAS [1]

Clientseitige UI-Generierung basiert auf der Idee von API-Antworten und zielt darauf ab, die Darstellung von UI-Elementen und Daten dynamisch auf der Client Seite zu generieren. Diese Prozedur lässt sich zum Beispiel über Bibliotheken wie *JSON Forms* ermöglichen, welche durch die Ausgabe der Daten in JSON in einem gewissen Schema vorliegen muss, so dass JSON Forms daraus automatisch Formulare auf der Client Seite generieren kann. Die Benutzeroberfläche wird dabei von drei Schemata definiert: einem Daten Schema, in welches die ausgegebenen Daten oder die zu sendenden Daten gespeichert werden, dem JSON-Schema, welches die Struktur der Benutzeroberfläche enthält (wie die zu generierenden Elemente, Eigenschaften und die dementsprechenden Typen) und dem UI-Schema, was bestimmt in welcher Ausrichtung und wie die Elemente angeordnet werden sollen. Alle drei Schemata werden zur Laufzeit vom Framework interpretiert und auf die entsprechenden UI-Komponenten abgebildet. In Abbildung 2 kann man das Zusammenspiel der Schemata sehen. Das Framework an sich bietet auch noch weitere Funktionalitäten wie verschiedene Eingabevalidierungen, beispielsweise mit Regex, dynamische Required-Abfragen, sobald sich eine Eingabe ändert und vieles mehr. JSON Forms zielt darauf ab, die Erstellung von komplexen Formularen zu vereinfachen, ganz nach dem Motto 'More forms. Less code.'. Das Framework bietet einfache Integrationen durch Pakete in React, Angular und Vue. Diese beinhalten auch gängige UI-Bibliotheken wie z.B. Material UI, ShadCn, und noch einige weitere, welche durch die Open-Source Community regelmäßig auf den neusten Stand gebracht und auch neue Features hinzugefügt werden. [2]

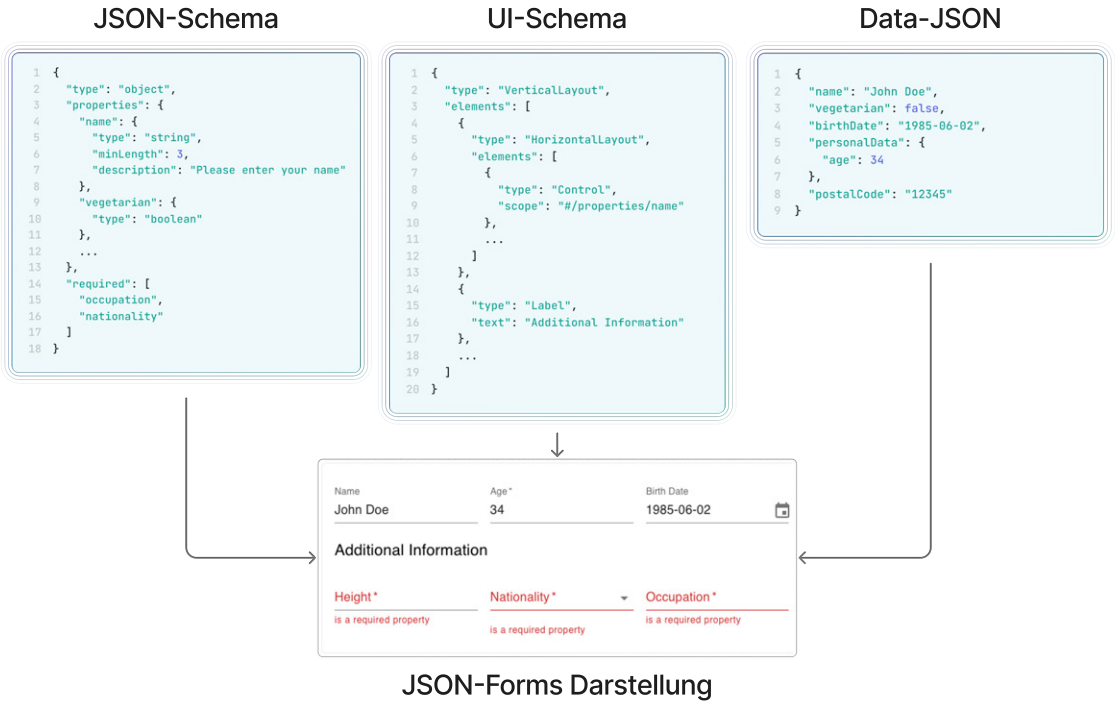


Abb. 2: Zusammensetzung von JSON Forms [1]

Anforderungen

Durch das Zusammenspiel mit HATEOAS soll es möglich sein, dem Client mitzuteilen, welche nächsten Aktionen möglich sind. Somit erfordert die Verwendung später in erster Linie keine großen Vorkenntnisse zum Aufbau der Anwendung bzw. der API. Wenn man dann auch noch die nötigen Formulare oder UI-Elemente für diese Aktionen beschreibt, kann der Client die Benutzeroberfläche dementsprechend dynamisch generieren.
Dadurch ergeben sich Vorteile wie die verbesserte Wartbarkeit, denn Änderungen am Backend, wie neue Integrationen, Ausgaben, usw. können nun problemlos geändert werden, ohne dass der Anwender zusätzliche Änderungen im Frontend vornehmen muss. Das führt unter anderem dazu, dass der Entwicklungsaufwand im Frontend reduziert wird. Weiterhin wird die API intuitiver in der Bedienbarkeit, denn durch das explizite Aufzeigen der nächsten verfügbaren Aktionen und deren Beschreibung kann die API vom Client besser verstanden werden.

Ausblick

Die Implementierung und Konzeption einer solchen API-Erweiterung könnte zukünftig weitere Perspektiven in der Anwendungsentwicklung eröffnen. Sie könnte durch die Verbindung beider Technologien, also HATEOAS und JSONForms wie Frontend und Backend miteinander agieren grundlegend verändern und auch neue Standards für die dynamische Softwareentwicklung setzen. Trotz der vielversprechenden Vorteile birgt die Umsetzung dieses Ansatzes auch Herausforderungen und Risiken, wie zum Beispiel die Komplexität der Daten, insbesondere bei sehr komplexen oder stark verschachtelten Datenstrukturen.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] JSON Forms. What is JSON Forms? https://jsonforms.io/docs/, 2021.

[3] Marc Müller. REST-API-Entwurfsmuster: HATEOAS, Resource Embedding. https://www.appleute.de/app-entwickler-bibliothek/rest-api-entwurfsmuster/, 2025.

[4] R. Padmanaban, M. Thirumaran, P. Anitha, and A. Moshika. Computability evaluation of RESTful API using Primitive Recursive Function. *Journal of King Saud University - Computer and Information Sciences*, 34:457–467, 2022.

Konzeptionierung und prototypische Implementierung für eine effiziente Nutzung zusätzlicher Fahrzeug-APIs für In-Car Apps

Fabian Etzler **Dennis Grewe**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Mercedes-Benz Tech Innovation GmbH, Gropiuspl. 10, 70563 Stuttgart

Einleitung

Die Digitalisierung hat das Automobil weit über seine ursprüngliche Funktion als Fortbewegungsmittel hinaus transformiert. Moderne Fahrzeuge sind mit hochentwickelten Displays und leistungsfähiger Hardware ausgestattet und eröffnen so eine neue Welt des Infotainments. Ein Beispiel hierfür ist die In-Car App "Tips", die den Kunden hilft, die volle Funktionalität ihres Fahrzeugs zu entdecken und zu nutzen. Diese App bietet maßgeschneiderte Erklärungen zu Fahrzeugfunktionen, illustriert durch Texte, Bilder und animierte How-to-Videos, und filtert Inhalte basierend auf der Fahrzeugkonfiguration [3].

Die zunehmende Komplexität der Fahrzeuge erfordert eine effiziente Verwaltung dieser Systeme und ihrer Interaktionen [2]. Um diese Herausforderung zu bewältigen, sind standardisierte Schnittstellen von entscheidender Bedeutung, da sie die Kommunikation zwischen den verschiedenen Komponenten vereinfachen. In diesem Kontext spielen Application Programming Interfaces (APIs) eine entscheidende Rolle. Sie ermöglichen eine reibungslose Interaktion zwischen Software und Fahrzeughardware und unterstützen somit die Steuerung und Erweiterung der Systeme.

Neben den Steuergeräten können auch In-Car Apps vereinzelt Fahrzeug-APIs nutzen, um mit der Hardware zu interagieren. Je nach Ausstattung und Softwareversion bieten diese APIs unterschiedliche Funktionalitäten. Mithilfe dieser APIs können In-Car Apps bspw. umfangreiche Fahrzeugdetails bereitstellen und Einstellungen wie bspw. den Fahrmodus oder die Abfahrtszeit setzen. Die vielfältigen Möglichkeiten, die die APIs bieten, ermöglichen es Entwicklern, innovative Anwendungen zu schaffen. Diese Anwendungen zielen darauf ab, den Komfort und die Sicherheit der Fahrerinnen und Fahrer weiter zu erhöhen.

Zielsetzung

Die erste API (im Folgenden API1 genannt) aus Abb. 1 ist bereits etabliert und bietet ein umfangreiches Funktionsangebot, das über einen Abstraktionsserver bereitgestellt wird. Der Server stellt eine Representational State Transfer (REST)-API zur Verfügung, um die Fahrzeug-APIs zentral zugänglich zu machen. API1 hat sich als zuverlässige Schnittstelle erwiesen, die eine Vielzahl von Fahrzeugfunktionen unterstützt und bereits erfolgreich in verschiedenen Anwendungen integriert wurde.

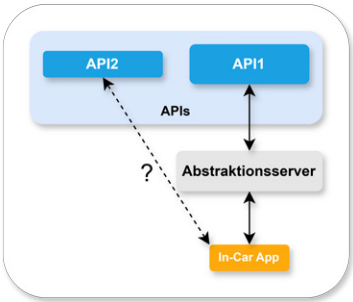


Abb. 1: Architektur ohne Anbindung [1]

Zusätzlich zu den Funktionen der API1 gibt es weitere APIs (im Folgenden API2 genannt), die für In-Car Apps interessante Funktionen bereitstellen. Diese APIs erweitern die bestehenden Möglichkeiten und bieten zusätzliche Funktionalitäten, die für die Entwicklung weiterer innovativer In-Car Apps erforderlich sind.

Das Ziel dieser Bachelorarbeit besteht darin, die zusätzlichen Funktionen der API2 für In-Car Apps zugänglich zu machen und zu evaluieren, wie diese neuen Schnittstellen genutzt werden können. Der Schwerpunkt liegt auf dem Testen und Evaluieren verschiedener Implementierungsansätze, um die am besten geeignete Lösung zu ermitteln. Dabei sollen

sowohl technische als auch nutzerbezogene Aspekte berücksichtigt werden.

Konzeptionierung

Im Rahmen der Konzeption wurden vier Lösungsansätze ausgearbeitet, um die API2 für In-Car Apps zugänglich zu machen. Im Rahmen einer umfassenden Analyse wurden die verschiedenen Ansätze hinsichtlich ihrer jeweiligen Vor- und Nachteile sorgfältig untersucht. Der Fokus lag darauf, die theoretische Machbarkeit und Benutzerfreundlichkeit jedes Ansatzes zu prüfen und diese anschließend anhand der zuvor aufgestellten Anforderungen zu bewerten.
Die Integration der API2 in den Abstraktionsserver der API1 wurde als die passendste Lösung betrachtet. Diese Lösung nutzt die vorhandene Infrastruktur und ermöglicht damit eine zentrale Nutzung der APIs. Die Integration wird in Abb. 2 dargestellt. Bei der Wahl dieser Lösung wurde bereits die zukünftige Skalierbarkeit und Erweiterbarkeit berücksichtigt.

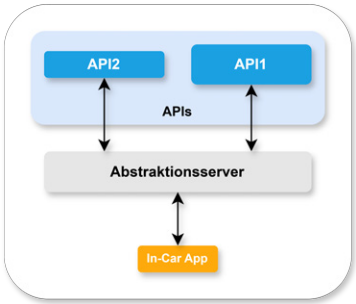


Abb. 2: Architektur mit Anbindung [1]

Umsetzung

Vor der Implementierung wurde beschlossen, sich auf zwei spezifische Signale der API2 zu beschränken: ein Signal aus der Kategorie ‚request/response' und ein Signal aus dem Bereich ‚publish/subscribe'. Diese zielgerichtete Auswahl ermöglichte eine fokussierte Herangehensweise an die Integration der API2.
Die Einarbeitung in den Abstraktionsserver zeigte jedoch schnell, dass dessen Komplexität eine erhebliche Herausforderung darstellte. Um dessen Struktur und Funktionsweise vollständig zu verstehen und die Implementierung der API2 vorzubereiten, war ein erheblicher Zeitaufwand erforderlich.
Nach dem erfolgreichen Überwinden anfänglicher Hindernisse konnten zusätzliche Endpunkte schließlich in den Server integriert und getestet werden. Die Tests haben die Funktionalität der neuen API-Endpunkte bestätigt und deren Fähigkeit, die gewünschten Daten bereitzustellen, validiert.
Zum Abschluss wurde ein vollständiger Systemtest durchgeführt, der die gesamte Kommunikation von der In-Car App bis zur API umfasste. Es waren auch Zeitmessungen geplant, um die Performance der API zu bewerten. Nach zahlreichen Versuchen musste der Systemtest leider abgebrochen werden, da keine Aussicht auf Erfolg bestand. Da die Zeitmessungen nicht über die In-Car App durchgeführt werden konnten, wurden alternative Messungen mit Curl vorgenommen.
Die Integration verlief damit grundsätzlich erfolgreich, jedoch führten die erforderlichen Konfigurationen für den Abstraktionsserver zu unerklärlichen Abstürzen der In-Car Apps.

Fazit

Die Integration der API2 in den Abstraktionsserver erwies sich als technisch anspruchsvoll, letztlich aber als machbar. Trotz der erfolgreichen Implementierung und Tests der neuen Endpunkte bestehen weiterhin Herausforderungen, insbesondere bei der Konfiguration des Servers für die Funktionalität der In-Car Apps. Die Arbeit zeigt, dass die Nutzung der vorhandenen Infrastruktur viele Vorteile bietet, jedoch auch spezifische Probleme mit sich bringt, die weiter untersucht werden müssen.
Um eine stabile und zuverlässige Nutzung der API2 zu gewährleisten, müssen zunächst die Ursachen für die Abstürze der In-Car Apps identifiziert und behoben werden. Darüber hinaus müssen die verbleibenden Signale der API implementiert werden, um den vollen Funktionsumfang der API2 verfügbar zu machen. Dadurch wird eine umfassende Nutzung für alle Entwickler ermöglicht.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Simon Fürst. Challenges in the design of automotive software. *IEEE*, 2010.

[3] Mercedes-Benz Tech Innovation. Unlock your car's full potential with the #tipsapp! https://www.linkedin.com/posts/mercedes-benz-tech-innovation_tipsapp-techinnovators-mercedesbenz-activity-7254425234743652354-9U0M, 2024.

Entwicklung einer Plausibiltätsprüfung basierend auf Grenzwertanalyse und Kurvendiskussion von definierten Bewegungen für eine pharmazeutische Abfüllmaschine

Joachim Faerber **Clemens Klöck**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Syntegon Technology GmbH, Crailsheim

Ausgangspunkt

Ausgangspunkt dieser Arbeit ist der Füllprozess von Abfüllmaschinen. Explizit soll es um die Füllbewegung gehen. Abhängig vom Maschinentyp kann diese unterschiedlich aussehen. Grundsätzlich besteht sie aus der Bewegung des Nadelbalkens und der Füllung des Objekts durch eine Pumpe. Im HMI (Human Machine Interface) können für den Prozess bestimmte Rezeptwerte zur Parametrierung (z.B. Zeit N01) des Prozesses angezeigt und verändert werden. In Abbildung 1 ist eine Visualisierung, wie sie auf dem HMI angezeigt wird, dargestellt.

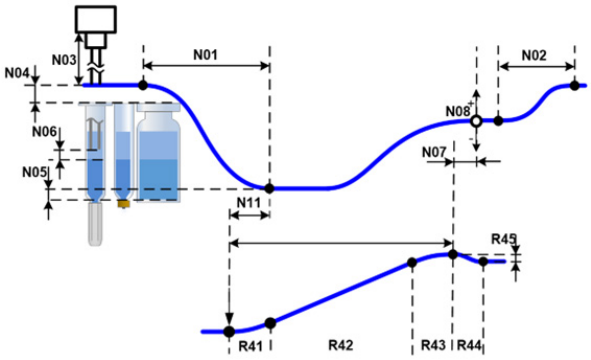


Abb. 1: HMI Visualisierung eines Füllprozesses mit Rezeptwerten [1]

Allerdings ist aus dieser Darstellung nicht ohne weiteres erkennbar, wie die parametrierte Bewegung nachher genau aussehen wird. Es ist also nicht sichtbar, ob bestimmte physikalische Grenzwerte, wie Position, Geschwindigkeit, Beschleunigung und Drehmoment innerhalb des aktiven Füllprozesses überschritten werden. Vor allem bei Tests mit verschiedenen Rezeptwerten, aber auch später beim Benutzen der Maschine kann dies leicht zu Fehlparametrierungen führen. Dies kann unter anderem auch kostspielige mechanische Schäden mit resultierenden Stillstandszeiten zur Folge haben.

Ziel der Arbeit

Um zuvor genannte Schäden vorzubeugen, Parametrierung zu vereinfachen und so den Zeitaufwand für diese zu verringern, soll eine sogenannte Plausibilitätsprüfung entwickelt werden. Diese Plausibilitätsprüfung soll Maximalwerte, die durch die Parametrierung der Bewegung entstehen, mit bestimmten Prozessgrenzen (z.B. Position, Geschwindigkeit, Beschleunigung und Drehmoment) vergleichen. Abbildung 2 veranschaulicht die Aufgabenstellung, deren Nutzen und die notwendigen Voraussetzungen, damit eine Plausibilitätsprüfung durchgeführt werden kann.

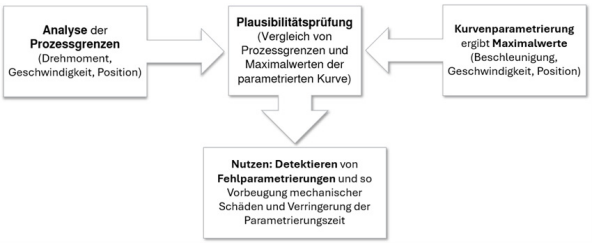


Abb. 2: Veranschaulichung der Aufgabenstellung [1]

Es müssen dazu in einem ersten Schritt der Füllprozess auf die relevanten Prozessgrenzen bzw. Ausgangsgrößen und die sich darauf auswirkenden Rezeptwerte bzw. Eingangsparameter analysiert werden. Notwendigerweise braucht man für den Vergleich auch die tatsächlichen Maximalwerte (z.B. für Position, Geschwindigkeit, Beschleunigung und Drehmoment), die sich aus einer parametrierten Bewegungskurve des Füllprozesses ergeben. Die Maximalwerte der Bewegung des Füllprozesses ergeben sich aus einem mathematischen Modell. Wenn diese theoretischen Grundlagen gelegt sind, kann in einem weiteren Teil das Programm für die Plausibilitätsprüfung in der konzeptioniert und später implementiert

werden. In einem letzten Schritt soll das Ergebnis in einem Integrationstest validiert werden.

Grenzwerte

Es sollen später verschieden Grenzen des Abfüllprozesses auf Plausibilität überprüft werden. Interessant sind dabei vor allem vier Größen: Position, Geschwindigkeit, Beschleunigung und Drehmoment. Da der Abfüllprozess nicht nur aus einem Bauteil besteht, kommen auch die Grenzwerte des Prozesses aus verschiedenen technischen Bereichen wie Softwaretechnik, Elektrotechnik, Mechanik und dem Prozess selbst. Dabei liegen die Grenzwerte nicht unbedingt in den geforderten Ausgangsgrößen. Aus diesen beiden Gründen sollen in diesem Abschnitt die vier Bereiche auf die daraus erschließbaren Grenzwertgrößen analysiert und eventuelle notwendige Umrechnungen vorgenommen werden. Ein Beispiel dafür ist die maximale Beschleunigung, die aus dem maximalen Drehmoment des Motors berechnet und später noch von einer rotatorischen, in eine lineare Beschleunigung umgerechnet werden muss, da die Ausgangsbewegung der Füllnadel eine lineare Auf -und Abbewegung ist (s.h. Abb. 1).

Ansatz des mathematischen Modells

Grundlage des mathematischen Modells zur Beschreibung des Bewegungsablaufs und Berechnung der Maximalwerte, die für die Plausibilitätsprüfung benötigt werden, ist die VDI-Richtlinie 2143 "Bewegungsgesetze für Kurvengetriebe". In dieser Richtlinie werden Grundlagen der Bestimmung von Bewegungsgesetzen für Kurvengetriebe vermittelt, die es dem Anwender ermöglichen sollen Bewegungspläne und Bewegungsdiagramme in Abhängigkeit geforderter Bewegungsaufgaben zur erstellen (vgl. VDI 2143, 2018, S.2 [2]). Wichtig für die Arbeit und spätere Programmierung sind dabei vor allem die Berechnung einzelner Bewegungsabschnitte in Abhängigkeit der Kombination der Bewegungsaufgaben und des Bewegungsgesetzes. Eine Beispielkombination dafür wäre das, in der Kurvenplanung weit verbreitete, Bewegungsgesetz eines Polynom 5. Grades und der Bewegungsaufgabenkombination Rast in Rast (s.h. Abb. 3).

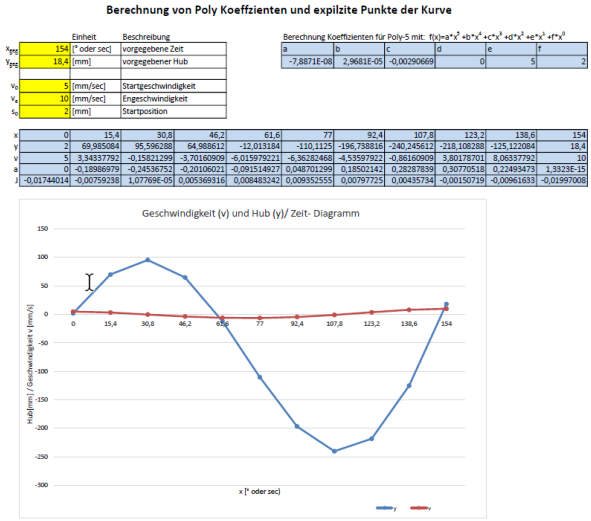


Abb. 3: Beispiel Excel- Berechnung eines Bewegungsgesetztes mit Form eines Polynom 5. Grades [1]

Da dieses Bewegungsgesetz häufig in Bewegungsabläufen Anwendung findet, ist es unter anderem auch eines der Bewegungsgesetze, dass später in der Programmierung des Funktionsbausteins der Plausibilitätsprüfung bevorzugt betrachtet werden wird.

Ausblick

Im weiteren Verlauf der Arbeit soll in einem ersten Schritt, anhand eines Beispiels für eine bestimmte Bewegungsaufgabe und Bewegungsgesetz einer bestimmten Füllstation und der Berechnung deren physikalischen Grenzwerte, ein Prototyp des Funktionsbausteins für die Plausibilitätsprüfung erstellt werden. Die Programmierung wird dabei in der Programmierumgebung IndraWorks der Firma Bosch Rexroth in der Programmiersprache ST (structured text) erfolgen. Im Anschluss wird der Prototyp des Funktionsbausteins iterativ an firmeninterne Standards angepasst und um weitere Features, wie andere Bewegungsgesetze -und aufgaben, weitere Grenzwertabfragen (z.B. minimale und maximale Position) erweitert. Dazu sollte der Baustein möglichst einfach erweiterbar und strukturiert programmiert werden. Zum Abschluss der Arbeit soll der Funktionsbaustein noch getestet werden.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Verein Deutscher Ingenieure VDI. *VDI 2143: Bewegungsgesetze für Kurvengetriebe*. Verein Deutscher Ingenieure (VDI), 1980.

Schnittstelle zur Automatisierung der Zugriffsverwaltung bzw. Berechtigungsvergabe innerhalb einer großen/diversen Softwarelandschaft

Mikail Anil Fidan **Thomas Rothermel**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Strabag, Stuttgart

Administration in der aktuellen SW-Landschaft

In unserer aktuellen Software-Landschaft werden Nutzer sowohl in Jira als auch in GitLab administriert, jedoch bisher ohne übergreifendes Konzept. Aktuell erfolgt die Verwaltung der Zugriffsrechte zwischen Jira und GitLab fast vollständig manuell und dezentral. IT-Administratoren vergeben, ändern oder entziehen Rechte einzeln auf Projekt- oder Repository-Ebene, meist über Tickets oder direkte Anfragen. Das bindet Zeit und Ressourcen und schafft Fehlerquellen: Ohne ein einheitliches Rollenkonzept bleiben beim Abteilungs- oder Rollenwechsel Berechtigungen oft unzutreffend bestehen und eröffnen Sicherheitslücken. Zudem ist nicht nachvollziehbar, wer wann welche Rechte erhalten hat, was für die ISO-27001-Zertifizierung kritisch ist. Diese verlangt ein lückenlos dokumentiertes Informationssicherheitsmanagement (ISMS) mit transparenter, überprüfbarer Rechtevergabe [2] – eine Anforderung, die der aktuelle manuelle Prozess nicht erfüllt.

Motivation und Bedeutung der Automatisierung

Die automatische Synchronisation der Berechtigungen zwischen Jira und GitLab reduziert den manuellen Administrationsaufwand drastisch. Anstelle vieler Klicks und starrer Handarbeitsprozesse erfolgt die Vergabe und Anpassung künftig automatisch, was eine deutliche Entlastung der IT-Abteilung bewirkt. Gleichzeitig sinkt das Fehlerrisiko, da Berechtigungen in Echtzeit aktualisiert und unbefugte Zugriffsdauern vermieden werden. Das Ergebnis sind schnellere Abläufe, weniger Fehler und frei werdende Kapazitäten. Darüber hinaus erfüllt der automatisierte Prozess die Anforderungen der ISO-27001, da alle Schritte lückenlos dokumentiert und nachprüfbar sind [2].

Forschungsfrage und Zielsetzung

Die Bachelorarbeit beleuchtet den Umfang der Entlastung von IT-Administratoren durch die Automatisierung der Schnittstelle zwischen Jira und GitLab – sowohl zeitlich als auch hinsichtlich der Anzahl der Klickschritte. Aktuell erfordert die manuelle Vergabe und Änderung von Berechtigungen einen ständigen Wechsel zwischen beiden Systemen sowie das Durchklicken zahlreicher Menüs und Tickets, was sowohl Zeit als auch Nerven bindet. Ziel der Untersuchung ist die Entwicklung und Implementierung eines funktionsfähigen Automatisierungstools zur Echtzeitsynchronisation von Berechtigungen, Minimierung von Fehlermöglichkeiten und automatischen Erstellung von Audit-Trails. Ein Vergleich der Bearbeitungszeiten und Klickschritte vor und nach Einführung des Tools ermöglicht eine quantifizierbare Darstellung der Effizienzgewinne und schafft die Grundlage für nachhaltige Ressourcen- und Sicherheitsoptimierung im Unternehmen.

Geplantes Vorgehen und Konzept

Derzeit existiert keine zentrale Automatisierung, die Jira-Rollen nahtlos in GitLab-Berechtigungen überführt. Vielmehr erfolgt die Vergabe manuell über die REST-API oder direkte Einstellungen in beiden Systemen, was zu Unübersichtlichkeit und hoher Fehleranfälligkeit führt. Der Jira2GitLab-Sync-Service übernimmt die Funktion eines Single Point of Truth: In Jira angelegte Rollen und Gruppen werden von einer Mapping-Engine automatisch den entsprechenden GitLab-Rechten zugewiesen. Jede Änderung – ob Anlegen, Ändern oder Löschen – löst umgehend Grant- bzw. Revoke-Operationen aus und beseitigt so mögliche Inkonsistenzen.

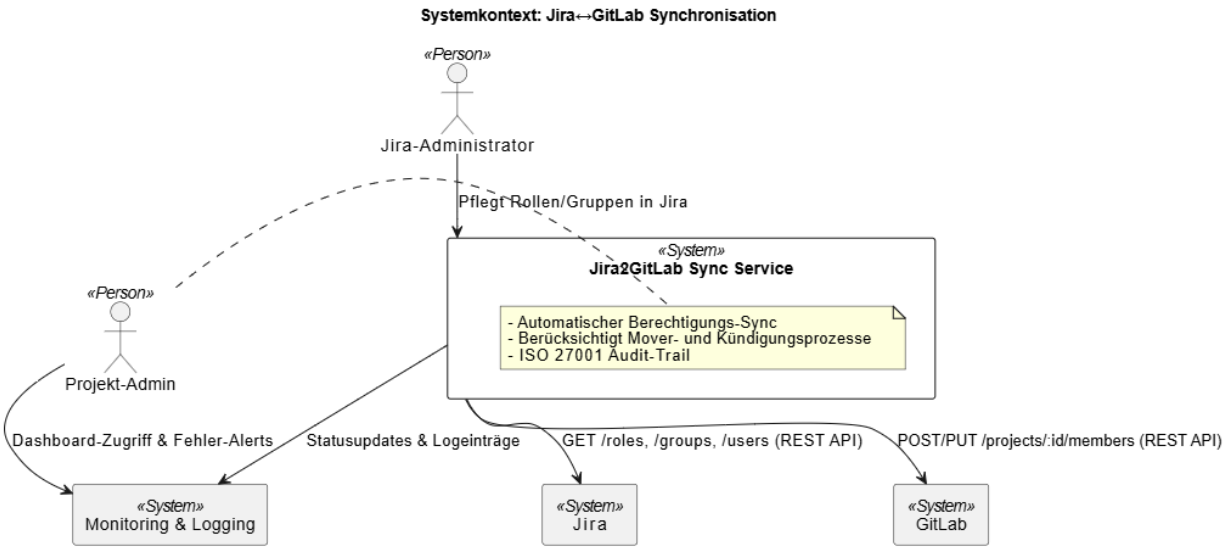


Abb. 1: Systemkontext Diagramm für die Software [1]

Der Jira-Administrator legt in Jira Rollen, Gruppen und Nutzer an bzw. ändert sie. Der Projekt-Admin überwacht den Synchronisationsstatus und reagiert auf Fehlermeldungen im Monitoring. Der Jira2GitLab Sync Service fungiert als Middleware, er:

- fragt periodisch neue oder geänderte Nutzer/Rollen aus Jira über die REST-API ab
- wendet ein exaktes Mapping auf die entsprechenden GitLab-Berechtigungen an
- automatisiert Mover- und Kündigungsprozesse
- erstellt ein ISO 27001-konformes Audit-Log

Im Kern steht ein genau definiertes Mapping: Jede Jira-Rolle erhält eine festgelegte Entsprechung in GitLab-Rechten. Dieses Mapping wird in einer Mapping-Engine hinterlegt und überwacht. Sobald in Jira eine Rolle geändert, hinzugefügt oder gelöscht wird, erkennt die Engine die Änderung und übersetzt sie automatisch in konkrete Grant- oder Revoke-Operationen in GitLab. Dadurch entfällt das manuelle Heraussuchen der passenden Berechtigung, und Inkonsistenzen werden eliminiert. Die REST-API-Kommunikation orientiert sich an etablierten Best Practices, um Skalierbarkeit, Sicherheit und Wartbarkeit zu gewährleisten. Die API arbeitet stateless, das heißt jede Anfrage enthält alle notwendigen Informationen, ohne Sessions auf dem Server [3]. HTTP-Methoden (GET, POST, PUT, DELETE) werden gemäß dem CRUD-Paradigma konsistent eingesetzt, um die Vorhersehbarkeit der Endpunkte sicherzustellen [3]. Für Authentifizierung und Autorisierung kommt ausschließlich HTTPS in Verbindung mit OAuth2.0 zum Einsatz, womit der Standard moderner REST-APIs umgesetzt wird [3]. Performance-Optimierungen erfolgen durch Caching gängiger Daten und Pagination bei umfangreichen Abfragen, um Latenzen zu reduzieren und Systemressourcen effizienter zu nutzen [3]. Die Versionierung der API (z. B. über URL-Muster /api/v1/…) ermöglicht die Einführung abwärtskompatibler Änderungen und eine flexible Weiterentwicklung des Dienstes [3]. Einheitliche Ressourcennamen und hierarchische Pfadstrukturen machen Endpunkte selbsterklärend und bilden die Beziehung zwischen Ressourcen ab [3]. Herausforderungen der Cache-Invalidierung werden durch gezielte Cache-Strategien und kontinuierliches Monitoring adressiert, um Datenfrische und Performance sicherzustellen [3].

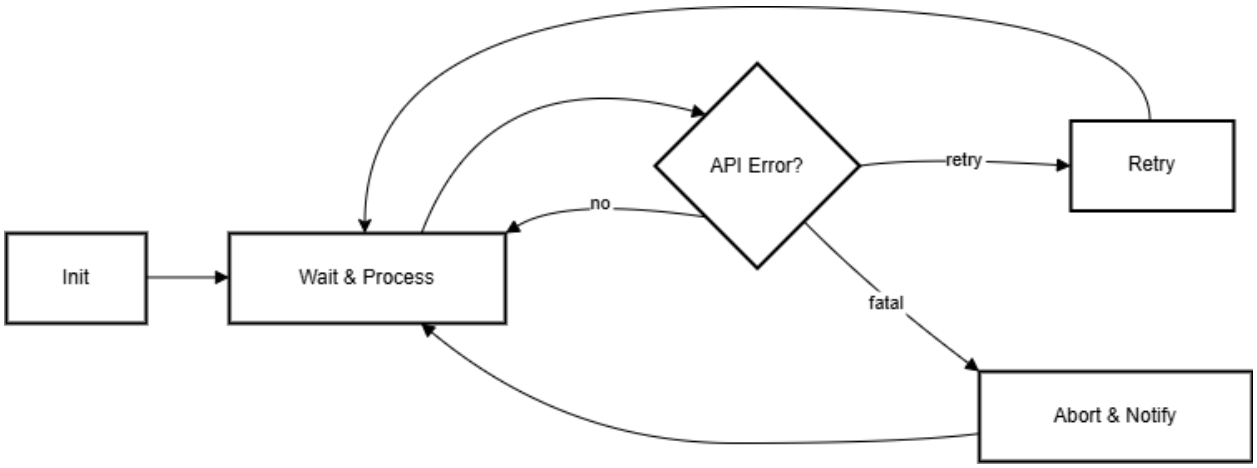


Abb. 2: Flowchart für die Software [1]

Abbildung 2 zeigt den Ablauf der Synchronisation:

- Init: Startet die Synchronisation.
- Wait & Process: Wartet auf Webhook-Events und verarbeitet sie.
- API Error?: Bei Fehlern wird geprüft, ob ein Retry möglich ist.
- retry: Ein erneuter Versuch wird unternommen.
- fatal: Der Batch wird abgebrochen und eine Benachrichtigung versendet.
- Falls kein Fehler vorliegt, kehrt der Ablauf zum Warten zurück.

Für Transparenz sorgt der Audit-Logger, der alle Änderungen ISO-27001-konform protokolliert, während das Monitoring Echtzeit-Alerts und Statusmeldungen liefert. Über eine einfache Admin-UI können Projekte und Rollen per Black-/Whitelist gesteuert und Mappings bei Bedarf angepasst werden. Dank skalierbarer Architektur, klarer Wartungs- und Support-Zuständigkeiten sowie umfassender Dokumentation bleibt die Rechteverwaltung langfristig sicher, effizient und vollständig nachvollziehbar.

Methodik zur zukünftigen Bewertung

Die Effizienzgewinne werden anhand zweier Kennzahlen ermittelt: Zeit pro Vorgang und Anzahl der Klickschritte. Hierzu erfolgt die Erfassung von Bearbeitungsdauer und Klicks für Standardaufgaben (Nutzeranlage, Rollenänderung, Berechtigungsentzug) sowohl im manuellen Verfahren als auch nach der Automatisierung – mittels Stoppuhr und Workflow-Tracking. Aus den Differenzen lassen sich konkrete Einsparwerte in Minuten und Klickschritten ableiten. Diese belastbaren Daten ermöglichen eine präzise Quantifizierung des Automatisierungseffekts und erlauben gezielte Optimierungen bereits während der Roll-out-Phase.

Erwarteter Ergebnisse

Es wird erwartet, dass die Automatisierung die Klickschritte deutlich reduziert und die Bearbeitungszeit für Berechtigungen spürbar verkürzt. Auf diese Weise werden IT-Ressourcen für wertschöpfende Aufgaben freigesetzt. Praxisbeispiele und Studien belegen, dass automatisierte Prozesse nicht nur die Effizienz steigern, sondern auch die Sicherheit erhöhen, indem Berechtigungen in Echtzeit synchronisiert und ungewollte Zugriffsfenster vermieden werden. Gleichzeitig verbessert sich die Nachvollziehbarkeit, da alle Änderungen lückenlos im Audit-Trail dokumentiert werden. Im ISO-27001-Kontext erleichtern solche strukturierten Schnittstellen Audits (vgl. ISO/IEC 27000:2018, S. 19 f.) und tragen langfristig zu stabiler Compliance und Sicherheit bei [2].

Fazit

Die automatisierte Schnittstelle entlastet IT-Administratoren deutlich, steigert nachhaltig die Effizienz im Berechtigungsmanagement, verwaltet Rechte in Echtzeit und erfüllt ISO 27001-Standards durch minimierte manuelle Fehler.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] International Organisation of Standadization ISO. *ISO/IEC 20007*. ISO, 5 edition, 2018.

[3] Gowda Priyank and Gowda Ashwath Narayana. best practices in rest api design for enhanced scalability and security. *Journal of Artificial Intelligence Machine Learning and Data Science*, 2024.

Konzeptionierung und Umsetzung von Echtzeitfunktionalitäten in einem modularen Testsystem auf der STM32H7 Plattform

Paul Freudenreich **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 Komplexität der heutigen eingebetteten Geräte wird das Testen und Evaluieren dieser Geräte auch stetig schwieriger. Die erhöhten Anforderungen an die Hardware hat auch zur Folge, dass sich Anforderungen an die Testplattformen verschärfen. Damit nun verschiedenste Hardware über eine Testplattform getestet werden kann, muss diese Plattform hoch flexibel sein und auch alle Anforderungen dieser spezifischen Hardware abdecken können. Einige Komponenten eines solchen Systems besitzen auch eine gewisse Zeitsensibilität und deshalb auch Echtzeitanforderungen. Damit die Qualität und Funktionalität dieser Komponente zuverlässig geprüft werden kann, muss die Testplattform echtzeitfähig sein. [2] Es gibt deshalb die Herausforderung, dass ein Testsystem möglichst hardwareunabhängig ist aber sogleich auch höchste zeitliche Vorhersehbarkeit bieten muss. Somit müssen solche Testsysteme auf einer hohen Abstrahierungseben arbeiten und dennoch schnell bzw. in Echtzeit reagieren.

Einführung und Zielsetzung

Im Rahmen dieser Arbeit wird die **Echtzeitfähigkeit** in ein bestehendes modulares Testsystem konzeptioniert, implementiert und letztendlich auch evaluiert. Dieses bestehende Testsystem erlaubt es mit Remote-Zugriff über Kommunikationsschnittstellen wie USB und Ethernet, Periphere einer Testhardware anzusteuern sowie auch externe Geräte zu testen. Um das Testen zu vereinfachen ist dieses Testsystem als eine abstrahierte Bibliothek aufgebaut und kann prinzipiell auf fast jeder Hardware verwendet werden. Dazu muss es nur eine Hardwarespezifische Firmware geben, welche diese Bibliothek implementiert und die benötigten Schnittstellen zur Verfügung stellt. Somit kann ein Tester auf einem Host-PC in Python, Tests schreiben. Diese Tests kommunizieren per **Remote-Zugriff** mit der Testhardware und es können direkt vom PC, Pins und Signale auf der Hardware manipuliert werden. Dabei muss der Tester nie verstehen hinter welcher Hardware sich ein Signal befindet. Es gibt sogenannte **Labels**, welche vom System bereitgestellt werden und zur Adressierung der Hardware dienen. Es muss beispielsweise nur das Label "LED_RED"verwendet werden, um eine LED anzusteuern. Das System findet den tatsächlichen Pin automatisch allein heraus. Ziel dieser Arbeit ist nun in dieses Testsystem möglichst genaue **Echtzeitfähigkeit** auf solchen Funktionalitäten aufzubauen. Dabei muss es für einen Tester möglich sein, über **Remote-Zugriff** eine Sequenz hochzuladen. Diese Sequenz wird daraufhin nach Fehlern überprüft. Ab diesem Punkt soll der Tester die Sequenz starten können und diese soll alle Aktionen in Echtzeit abarbeiten. Als Hardwareplattform soll dabei die schon implementiert **RP2040** Plattform verwendet werden, sowie auch eine neue Plattform für die **STM32H7** Familie implementiert werden.

Konzept und Implementierung der Echtzeitsequenz

Um eine Sequenz von Aktionen innerhalb des Testsystems in ein für den Tester verständliches Format zu bringen, wird ein Textformat verwendet. Dabei wird eine Sequenz bzw. Programm in einer Textdatei geschrieben und zu einer einfach dekodierbaren Binärdatei konvertiert. Es ist zu beachten, dass diese Testsystem auf einer abstrahierten/hardwareunabhängigen Softwareebene agiert. Somit kann diese Binärdatei nicht direkt auf dem Mikrocontroller ausgeführt werden. Da nicht klar ist, auf welcher Maschine das Programm laufen soll, wird die Binärdatei von einem virtuellen Prozessor dekodiert und ausgeführt. Es gibt somit einmal eine **Textdatei**, welche die Sequenz

im Klartext enthält und nebenbei eine **Binärdatei**, welche schnelle Ausführung innerhalb des virtuellen Prozessors erlaubt. Eine zusätzliche Anforderung an diese Echtzeitfähigkeit ist, dass diese Sequenz auch anhand von Bedingungen, gewisse Aktionen wiederholen oder überspringen können muss. Aufgrund dieser zusätzlichen Anforderung wurde das Konzept dieser Sequenz als ein assemblerartiges Programm implementiert. Einer der Vorteile von Assembly ist, dass Bedingungen über Verzweigungsinstruktionen einfach implementiert werden können. Außerdem ist Assembly bekannt und dient als eine Vertrautheit für Nutzer dieser Implementierung. Somit muss keine komplett neue Syntax erfunden werden, welche die Nutzer lernen müssen.

Instruktionssatz

Es wird ein kleiner Instruktionssatz verwendet, welcher stark an die ARM Assembly Syntax anlehnt. Es gibt Instruktionen wie **MOV, ADD, INC,** usw. zur Manipulierung von Registern. [4]
Dabei gibt es jedoch auch für das Testsystem spezifische Instruktionen wie beispielsweise **SETB „RED_LED“, 1** was eine LED aufleuchten lassen würde. Hierbei werden die Labels des Testsystems wieder verwendet und somit erbt die Echtzeitimplementierung ebenfalls die Vorteile der Abstrahierung des Testsystems. Die Besonderheiten dieser Sprache sind somit folgende:

- Die Sprache ist einfach verständlich für Tester aufgrund der Vertrautheit mit **Assembly-Syntax**
- Es existieren Sonderinstruktionen, welche anhand von **Labels** die Periphere der Hardware abstrahieren
- Jede Instruktion kann mit einem **Zeitstempel** versehen werden

Der Zeitstempel einer Instruktion gibt dem Prozessor vor, ab welchem Zeitpunkt diese Aktion ausgeführt werden soll. Somit ist die Echtzeitfunktion direkt im Klartext ablesbar. Es wird ebenfalls ein Compiler und Disassembler implementiert. Wie in Abbildung 1 dargestellt, kodiert der Compiler die Textdatei in eine Binärdatei. Der Disassembler kann diese Binärdatei auch wieder zu Debuggingzwecken dekodieren und im Klartext anzeigen. [5]

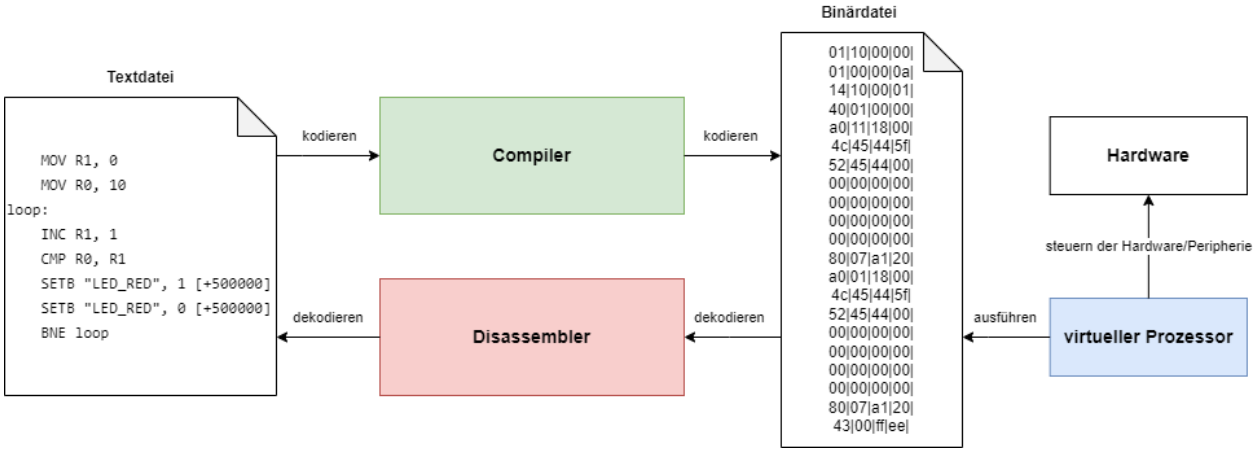


Abb. 1: Echtzeitumgebung innerhalb des Testsystems [1]

Virtueller Echtzeitprozessor

Um den abstrahierten Binärcode auszuführen, wird in das Testsystem ein **virtueller Prozessor** eingebaut, um die Instruktionen zu dekodieren und schnellstmöglich bzw. zum richtigen Zeitpunkt weitere benötigte Funktionalitäten des Systems abzurufen. Außerdem besitzt dieser virtuelle Prozessor folgende Komponente: [3]

- 16 Register mit 32 Bit Größe
- Ein „zero“ und „signed“ flag
- Programmzähler
- Statusflags wie „running“, „terminated“, „breaking“, ...
- Laufzeitzähler in Mikrosekunden
- Puffer des laufenden Programms in RAM
- Zugriff auf prinzipiell alle internen Funktionalitäten des Testsystems
- Sonstige weniger relevante Komponente

Da der virtuelle Prozessor nicht 100 % der Prozessorlaufzeit erhält, wird dieser mit einem vordefinierten Takt von 100 Mikrosekunden aufgerufen. Somit ist eine

kleinste Zeit von 100 Mikrosekunden zwischen zwei Instruktionen möglich. Der grobe Ablauf innerhalb der Prozessorroutine ist in Abbildung 2 dargestellt.

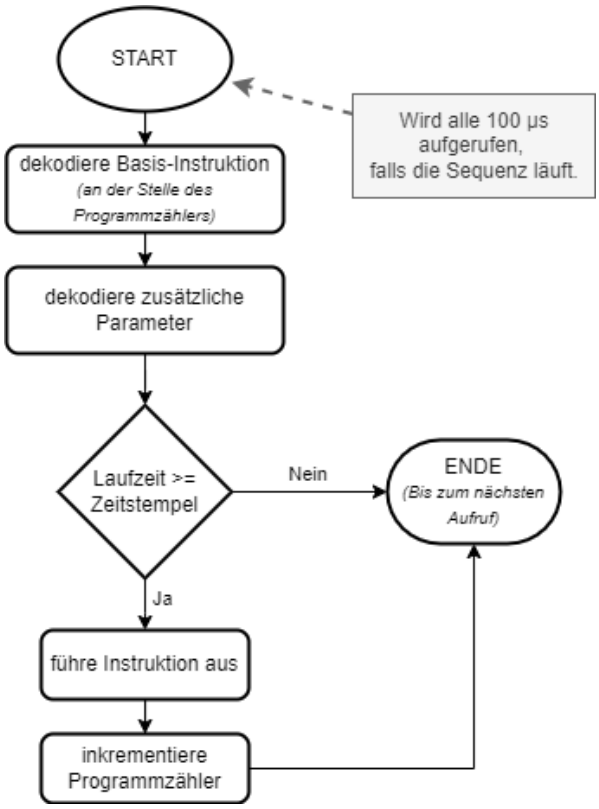


Abb. 2: PAP des virtuellen Echtzeitprozessors [1]

Evaluierung und Benchmarking

Um die Echtzeitfähigkeit des Systems nun zu evaluieren, gibt es viele Möglichkeiten. Es wurde als einfaches Benchmarking die GPIO-Toggle Funktionalität verwendet. Hierbei wird in der Instruktion einfach ein spezifizierter GPIO-Pin getoggelt. Da der Prozessor in einem 100 Mikrosekunden Takt agiert, wird eine Periode von 200 Mikrosekunden erwartet. Dies ist in Abbildung 3 im unteren Histogramm zu sehen. In der Abbildung ist der **Durchschnitt** mit **„mean“** und die **Standardabweichung** mit **„std“** dargestellt. Das obere Histogramm stellt das Verhalten bei demselben Programm ohne Echtzeit dar. Es kann also die Aussage gemacht werden, dass sich das System auf Hinsicht der Echtzeitfähigkeit verbessert hat. Die Präzision liegt hier bei einer Abweichung von 1 bis 10 Mikrosekunden, was die Systemanforderungen befriedigt.

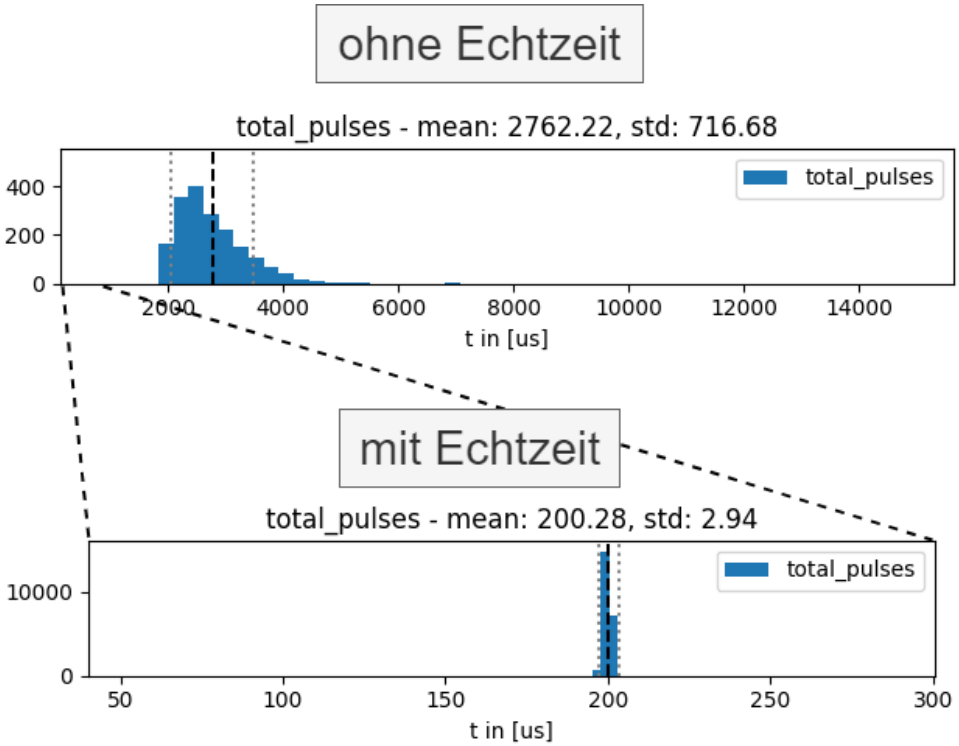


Abb. 3: Benchmarks der GPIO-Toggle Funktionalität [1]

Ausblick

Die Echtzeitfunktionalität bietet neues Potenzial für die Testplattform und ist in vielen Weisen auch ausbaubar. Es gibt Möglichkeiten die Implementierung zu optimieren und neue Funktionalitäten in den Instruktionssatz hinzuzufügen. Beispielsweise wäre hier ein Stack in den virtuellen Prozessor einbaubar, um große Datenmengen abzuspeichern. Außerdem kann das Modul aufgrund der Hardwareunabhängigkeit in neue Hardware problemlos implementiert werden. Es kann somit in der Zukunft weiteren Testern bei jeglichen Echtzeitkomponenten dienen.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Wolfgang Halang and Herwig Unger. *Industrie 4.0 und Echtzeit*. Springer Vieweg, 2014.

[3] AWS Inc. Was sind die Komponenten einer CPU. https://aws.amazon.com/de/what-is/cpu/, 2024.

[4] ARM Limited. Writing ARM Assembly Language. https://developer.arm.com/documentation/dui0473/m/writing-arm-assembly-language?lang=en, 2010.

[5] Manish Varshney and Vivek Sharma. *Design and Implementation of Compiler*. New Delhi: New Age International, 2009.

Entwicklung eines sprachgesteuerten KI-Assistenzsystems für Menschen mit Sehbehinderung mittels der tragbaren Sensorbrille Aria

Mateusz Frydryszak **Markus Enzweiler**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

Spracherkennung ist längst keine Science-Fiction mehr, sondern ein fester Bestandteil moderner Technologien. Sie ermöglicht die Umwandlung gesprochener Sprache in Text oder Steuerbefehle und findet breite Anwendung in digitalen Assistenten wie Siri und Alexa, in automatisierten Callcenter-Systemen oder in der Sprachsteuerung von Fahrzeugen. Im Rahmen dieser Arbeit wird ein sprachgesteuertes KI-Assistenzsystem auf Basis der tragbaren Sensorbrille Aria [2] von Meta entwickelt und evaluiert. Das System ermöglicht es, visuelle Bildaufnahmen aus der Perspektive der nutzenden Person an ein Vision-Language-Model (VLM) zu übertragen, das daraufhin kontextbezogene Beschreibungen oder Rückmeldungen über das Umfeld liefert. Im Mittelpunkt steht die sprachbasierte, händefreie Interaktion mit dem System, die insbesondere auf die Anforderungen sehbehinderter Personen ausgerichtet ist.

Zielsetzung

Schwerpunkt dieser Arbeit ist die Konzeption, prototypische Entwicklung und Evaluierung eines sprachgesteuerten KI-Assistenzsystems für sehbehinderte Menschen unter Verwendung der tragbaren Sensorbrille Aria von Meta. Ziel des Systems ist es, den Nutzenden zu ermöglichen, durch sprachliche Abfragen Informationen zum Umfeld zu erhalten. Insbesondere sehbehinderten Personen soll dadurch ein Assistenzsystem zur Verfügung gestellt werden, das mithilfe von Kamerabildern aus der Augenperspektive (First-Person-View) als visueller Ersatz dient und kontextbezogene sowie aufklärende Rückmeldungen liefert. Darüber hinaus werden innerhalb dieser Arbeit relevante Methoden wie Keyword-Spotting (KWS) und transkribierende Sprachverarbeitung untersucht.

Ansatz

Das Gesamtsystem setzt sich aus vier interagierenden Funktionsblöcken zusammen:

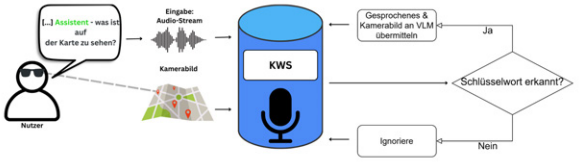


Abb. 1: Systemdiagramm [1]

- Keyword-Spotting (KWS): Ein Worterkennungs-System lauscht kontinuierlich auf ein vordefiniertes Aktivierungswort. Wird dieses erkannt, startet es das nachgelagerte Transkription-System
- Transkription: Nach der Aktivierung zeichnet das System die nachfolgenden gesprochenen Äußerungen des Nutzers auf und wandelt sie in Text um. Das zum Zeitpunkt der Aktivierung aufgenommene Kamerabild sowie die transkribierte Spracheingabe werden – gemeinsam mit nutzerspezifischen Konfigurationsparametern – als Prompt für das VLM verwendet
- VLM: Das VLM verarbeitet den textuellen Prompt gemeinsam mit dem aktuellen Kamerabild als Eingabe und erzeugt eine prägnante Antwort im Stil eines digitalen Assistenten
- Text-to-Speech (TTS): Die vom VLM erzeugte Antwort wird abschließend durch ein TTS-Modul in gesprochene Sprache umgewandelt und dem Nutzer auditiv ausgegeben

Bewertungsmetriken

Ein zentraler Aspekt des Assistenzsystems ist die Sicherstellung einer zuverlässigen Sprachverarbeitung mit möglichst geringer Latenz. Im Rahmen dieser Arbeit beschränkt sich die Evaluierung auf die Open-Source-Spracherkennungs-Frameworks Vosk, Faster-Whisper und Porcupine. Zunächst ist es notwendig, die Technologien anhand eines geeigneten Datensatzes zu testen, um sie anschließend anhand der folgenden Metriken zu bewerten:

- durchschnittliche Latenzzeit (in Sekunden)
- Erkennungsgenauigkeit
- F1-Score

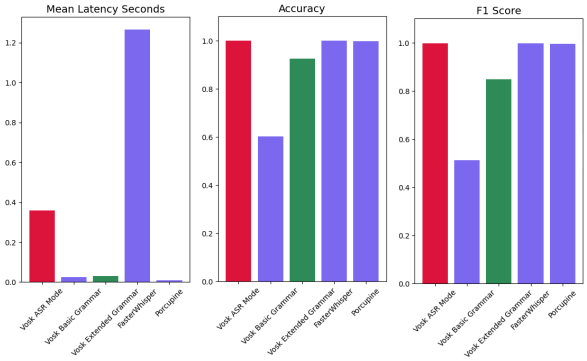


Abb. 2: Gemessene Metriken in einem KWS Use Case [1]

Die Messergebnisse weisen auf eine notwendige Abwägung zwischen Genauigkeit und Latenzzeit bei der Auswahl der geeigneten Technologie hin. Alle getesteten Frameworks erzielten einen F1-Score von mindestens 0,8, was auf ein ausgewogenes Verhältnis von Precision und Recall schließen lässt. In Kombination mit den Genauigkeitswerten deutet dies auf eine zuverlässige Schlüsselworterkennung hin. Getestet wurde mit dem Open-Source Datensatz Wake-Word-Benchmark von Picovoice.

Ausblick

Diese Arbeit zeigt, wie tragbare Sensorbrillen wie Aria einen bedeutenden Beitrag zur Barrierefreiheit leisten und insbesondere Menschen mit Sehbehinderung beim Navigieren in urbanen Umgebungen unterstützen können. Angesichts der schnellen Weiterentwicklung von VLMs stellt sich die Frage, in welchem Maße künstliche Intelligenz menschliche Fähigkeiten sinnvoll ergänzen und insbesondere Menschen mit Behinderungen entlasten kann. Darüber hinaus wirft die Arbeit grundlegende Fragen dazu auf, wie technologische Systeme künftig noch intuitiver und vollständig freihändig bedient werden können.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Platforms Meta. A wearable computer in a glasses form-factor. https://www.projectaria.com/glasses, 2025.

Concept and implementation of a TSN network for evaluation of IEEE/IEC 60802 technologies and OPC UAFX applications

Felix Geiger **Michael Scharf**

Department of Computer Science and Engineering, Esslingen University

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

Introduction

For many years, industrial automation has favored the integration of classic bus topologies with industrial Ethernet. The problem with the current Ethernet standard is that it cannot deliver the bounded latency, reliability and time-synchronization required for many industrial automation and robotics applications. Therefore, there are many different standards for Industrial Ethernet communication, such as PROFINET, EtherCAT, EtherNet/IP, and many more. These are based on the current Ethernet standard defined in IEEE 802.3 and extensions for the respective application. The problem that can arise for industrial networks is the partial lack of interoperability of the Industrial Ethernet protocols, which requires additional hardware such as gateways to enable communication between different network segments. One approach to solve this is currently defined in IEC/IEEE 60802. It is using Time-Sensitive Networking (TSN), adapted to industrial requirements, with the purpose of creating a standardized version of deterministic real-time Ethernet. TSN originates in the Audio and Video Bridging industry and focuses on extensions to the IEEE 802.1Q - Bridges and Bridged Networks Standard to realize bounded latency, high availability, and reliability. TSN for Industrial Automation (TSN-IA) is a joint project of the Institute of Electrical and Electronics Engineers (IEEE) and the International Electrotechnical Commission (IEC).

TSN-IA

TSN-IA standardizes a set of TSN features to ensure deterministic Ethernet communication with bounded latency, high reliability, and precise time synchronization. Like TSN, it builds on the IEEE 802.1Q Standard, which defines the Virtual Local Area Network (VLAN) alongside various other features. Using the VLAN Header enables setting different priorities and VLAN identifiers, providing the ability to identify different traffic classes [4]. Using this traffic class assignment enables the use of the Quality of Service mechanisms necessary to prioritize time-sensitive traffic over best-effort communications. The simplest approach is a strict priority queuing algorithm which will queue the traffic based on the priority [4]. Another approach originally defined in IEEE 802.1Qbv is to use time-aware scheduling [2]. The approach defined in 802.1Qbv is also called Enhancements for Scheduled Traffic (EST) or Time-Aware Shaper (TAS).

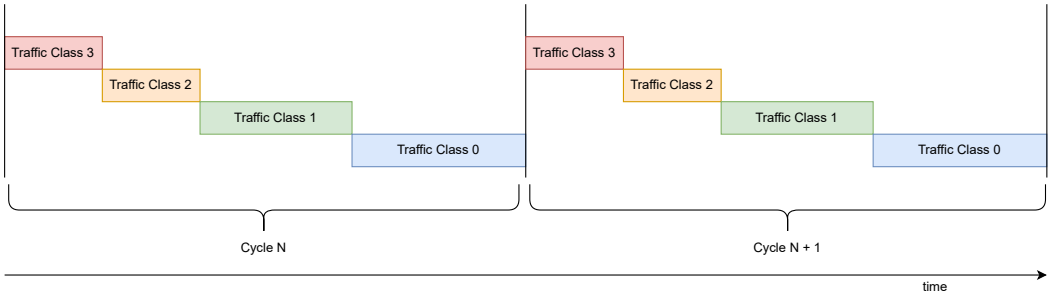


Fig. 1: Time-Aware Shaper example [7]

Figure 1 illustrates an example of how the TAS works. Basically, it creates transmission windows for different traffic classes by disabling transmission for other traffic classes in those time windows. The TAS is applied cyclically and will have the same order of transmission windows in each cycle. Another tool included in TSN-IA is Frame Preemption (FP). It is initially defined in IEEE 802.1Qbu and 802.3br. As shown in figure 2, frame preemption allows express packets to intersperse preemptable traffic. This is done by fragmenting the preemptable packet currently on the wire and sending the express packet. Afterwards, the rest of the preemptable packet can be transmitted, ultimately creating the egress traffic as illustrated.

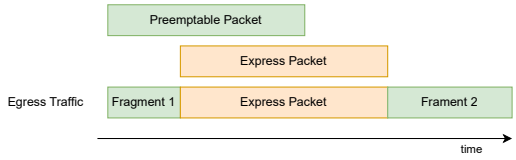


Fig. 2: Frame Preemption example [7]

To achieve the time synchronization necessary for features like TAS, TSN-IA uses a Precision Time Protocol (PTP) profile called generalized PTP (gPTP). The gPTP is defined in IEEE 802.1AS. Basically, a network using gPTP is synchronized by a determined Grandmaster (GM) Clock with a precise clock and time-source [3]. gPTP uses different methods to approximate the delay in-between two instances and applies this delay to a timestamp sent by the GM Clock to synchronize the local device. There are further features defined for Resource Management and Reliability. Currently, those features are not supported by most of the hardware.

OPC UAFX

As TSN only includes standards working on the second layer of the Open Systems Interconnection (OSI) model, the rest of the communication stack is undefined and various different communication models could be used. With Open Platform Communication (OPC) Unified Architecture (UA), the OPC Foundation offers a standardized communication model that is developed in close cooperation with the industry. To extend OPC UA to the field, the OPC Foundation has launched the Field eXchange (UAFX) specification initiative. Alongside the OPC's PubSub model, which was first introduced in 2018, it specifies the usage of TSN [5]. PubSub uses one-to-many communication by enabling a publisher to send to multiple subscribers [6]. Furthermore, there is an ongoing OpenSource project implementing the OPC UA specifications called the open62541 stack.

Hard- and Software setup

Implementing TSN-IA requires hardware that explicitly supports real-time networking features. This includes precise timestamping, hardware queues, and support for scheduling algorithms like Time-Aware Shaping and Frame Preemption. Software implementations alone cannot satisfy the deterministic requirements of TSN communication.
The setup used in this project involves an Intel I226 NIC, paired with a Linux system running the PREEMPT_RT real-time kernel patch. This provides a soft real-time capable platform suitable for evaluating TSN features. Intel I226 controllers offer gPTP support, time-aware scheduling, and multiple hardware queues, making them viable for TSN applications [1]. Configuration is performed using open-source tools such as linuxptp, the iproute2 package (network stack configuration) and the ethtool (for driver configuration). Other tools like MoonGen, iperf3 and open62541 are used for traffic generation and OPC UA integration. Alternatives include the RealTime-HAT from InnoRoute for Raspberry Pi systems. It includes a basic TSN software package and optional licensing for advanced features. More specialized implementations can be realized using TSN IP Cores, which allow FPGA-based integration of TSN features for fully customized TSN endpoints. For switching devices, switches from Belden and Phoenix Contact will be used.

Implementation and Outlook

The next step in this work involves the deployment of various network test setups to evaluate the performance benefits of TSN-IA features compared to standard Ethernet. Early experiments using Intel I226 NICs and a correctly configured Linux environment with the PREEMPT_RT patch indicate promising support for features such as time synchronization and scheduling.

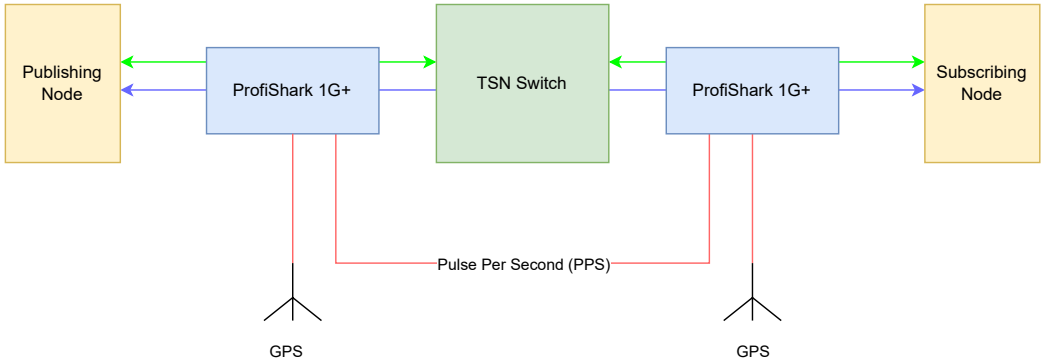


Fig. 3: Simple network setup for testing [7]

The evaluation will be based on multiple test topologies and will focus on combinations of different traffic scheduling algorithms and Frame Preemption. Key performance indicators (KPIs) include end-to-end latency, jitter, and packet loss under different traffic conditions and scheduling algorithms. An example topology is shown in figure 3. This basic setup uses two Linux-based nodes for traffic generation in combination with a TSN-capable switch. The green line illustrates communication between switches and node devices; an example of such communication would be time-synchronization using gPTP. The purple-colored line illustrates end-to-end traffic between the two node devices. In this basic setup, there would be two different applications running on the nodes: iperf3 and open62541 PubSub. Using iperf3 as a simulation for low-priority traffic and the PubSub communication for higher priority, simulating a best-effort and a high-priority application running in an industrial environment. The KPIs shall be evaluated separately and improvements shall be analyzed. For traffic capturing, the ProfiShark 1G+ devices are used. The synchronization of the capturing devices is achieved using the Global Positioning System (GPS) and an interconnected Pulse Per Second (PPS) signal.

References and figures

[1] Intel Corporation. Product Brief Intel Ethernet Controller I225 & I226. https://www.intel.com/content/www/us/en/content-details/621753/intel-ethernet-controller-i225-i226-product-brief.html, 2024.

[2] Institute of Electrical Engineers and Electronics. IEEE Standard for Local and metropolitan area networks – Bridges and Bridged Networks - Amendment 25: Enhancements for Scheduled Traffic. *IEEE Std 802.1Qbv-2015*, 2015.

[3] Institute of Electrical Engineers and Electronics. IEEE Standard for Local and Metropolitan Area Networks–Timing and Synchronization for Time-Sensitive Applications. *IEEE Std 802.1AS-2020 (Revision of IEEE Std 802.1AS-2011)*, 2020.

[4] Institute of Electrical Engineers and Electronics. IEEE Standard for Local and Metropolitan Area Networks–Bridges and Bridged Networks. *IEEE Std 802.1Q-2022*, 2022.

[5] OPC Foundation. Extending OPC UA to the field: OPC UA for Field eXchange (FX). *Extending OPC UA to the field: OPC UA for Field eXchange (FX)*, 2021.

[6] OPC Foundation. UA Part 14: PubSub. https://reference.opcfoundation.org/Core/Part14/v105/docs/, 2024.

[7] Own representation.

Formulation of a validation methodology for machine learning-enabled partially automated driving systems.

Yael Glaser **Markus Enzweiler**

Department of Computer Science and Engineering, Esslingen University

Work carried out at Mercedes-Benz AG, Böblingen

Introduction

Reliability, robustness, and safety are cornerstone requirements in the development of Advanced Driver Assistance Systems (ADAS). As artificial intelligence (AI) transforms the technological landscape, adapting development approaches is essential to address the unique challenges it introduces. "ADAS relies heavily on deep learning to analyze and interpret large amounts of data obtained from a wide range of sensors" [1]. In recent years, AI has surged to the forefront of innovation, offering immense opportunities for the automotive industry—an industry often slow to adopt major technological shifts. However, this advancement comes with significant challenges and responsibilities, particularly in safety-critical applications where failures can have severe consequences, such as accidents or diminished trust in automation. Until recently, automotive software development and validation focused on deterministic systems, where predictable behavior simplified quality assurance and error detection, guided by standards like the International Standard for Organization (ISO) 26262 for functional safety in road vehicles. The integration of AI, both in development and at runtime, disrupts this paradigm by introducing unpredictable system behavior due to the non-deterministic nature of machine learning (ML) models. In safety-critical contexts, this unpredictability heightens the risk of unforeseen failures. This thesis addresses these challenges by exploring the opportunities and limitations of AI in automated driving systems, focusing on enhancing existing validation practices.

Challenges of AI in ADAS

AI, particularly machine learning (ML), allows ADAS to process vast amounts of sensor data—such as camera and radar inputs—to detect obstacles and make driving decisions. However, unlike traditional automotive systems, which behave predictably, AI introduces uncertainty. ML models can struggle in unfamiliar situations, such as poor weather or complex urban environments, potentially leading to accidents or reduced driver trust. Additionally, many ADAS functions share the same ML models; for example, Emergency Steering Support (ESS) and Automatic Emergency Braking (AEB) often rely on a single object detection model. If this model fails, it can affect multiple safety features at once, making validation critical. Testing shared models is particularly difficult because a flaw in one function, like ESS, could compromise another, such as AEB, creating a ripple effect across the system. This interconnectedness adds pressure to ensure every scenario is accounted for, especially in high-stakes situations.

Goal

The primary goal of this thesis is to develop a reliable and comprehensive concept for validating automated driving functions that rely on machine learning modules, enhancing existing testing strategies to address the unique challenges of AI-driven systems. This concept will build on established practices, which are guided by safety standards like ISO 26262 for functional safety road vehicles, by incorporating advanced methods to tackle ML-specific risks. The resulting concept aims to enhance test coverage, improve validation efficiency, and ensure that AI-based functions meet stringent safety requirements. This approach will be applied to the Emergency Steering Support (ESS) function, with a focus on verification of shared ML models, such as the object detection model used for perception in both ESS and Automatic Emergency Braking (AEB). Since these shared models support multiple functions, specific test cases will be designed and executed to assess their performance across diverse scenarios, ensuring consistency and reliability. The results will be compared against theoretical expectations to confirm the concept's validity. In parallel, interviews with industry experts will provide practical insights to refine the concept. By blending rigorous development with real-world feedback, this thesis aims to deliver

a validation strategy that significantly improves the quality and reliability of testing for AI-based ADAS functions, complementing existing standards-driven approaches. This strategy could pave the way for safer, more reliable ADAS features in future vehicles. It may also help automakers build greater trust with drivers who rely on these systems daily.

Testing Principles

Testing is a fundamental aspect of the development process, ensuring the trustability, reliability, and quality of a product. However, testing is inherently time- and resource-intensive, making an effective and efficient testing strategy essential for project success. Testing occurs at multiple levels throughout the development lifecycle, as illustrated in the V-model in Figure 1, where each level—from unit testing to system validation—verifies different aspects of the system. A key principle emphasized across testing methodologies is the importance of independent testing, where tests are conducted by individuals or teams separate from the developers to ensure objectivity and reduce bias. Testing principles, which define the scope and approach to testing, form the foundation for all verification and validation procedures in the development process.

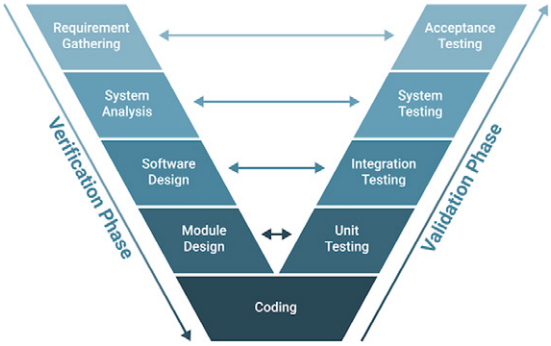


Fig. 1: V-Model [2]

A key concept in optimizing testing efforts is the testing pyramid, which visualizes the ideal distribution of testing activities across the V-model's levels as seen in Figure 2. "The "Test Pyramid" is a metaphor that tells us to group software tests into buckets of different granularity." [4] It suggests that the majority of testing effort should be concentrated at the lower levels of the V-model, such as unit and integration testing, where tests are detailed, precise, and extensive. At these levels, individual components (e.g., software modules, hardware units) and their interactions are thoroughly verified, catching defects early and reducing the likelihood of issues propagating to higher levels. As one progresses up the V-model to system testing and acceptance testing, the scope of testing broadens, but the effort and number of test cases should ideally decrease. This is because lower-level testing ensures that most defects are resolved early, allowing higher-level testing—like system testing—to focus on verifying overall functionality and compliance with system requirements, rather than uncovering new defects.

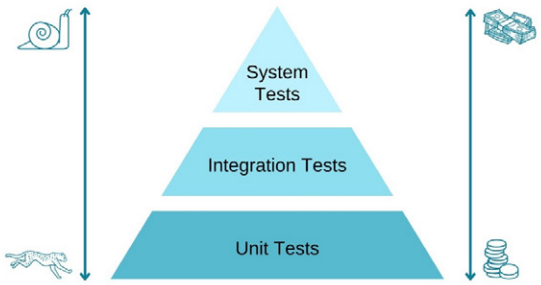


Fig. 2: Test Pyramid [3]

Outlook

As machine learning models become more complex and widely shared across ADAS functions, scalable and scenario-driven validation strategies will be essential. Future research should explore automated and simulation-based testing methods, supported by real-world data, to improve efficiency and robustness. Moreover, closer collaboration between industry and regulatory bodies will be crucial in shaping guidelines that ensure the safe deployment of AI in safety-critical applications. The validation concept developed in this thesis, focused on ESS, could serve as a starting point for these collaborative efforts, ensuring AI's safe integration into tomorrow's vehicles.

References and figures

[1] Ambuj Nandanwar. Artificial Intelligence (AI) utilizing deep learning techniques to enhance ADAS. https://www.design-reuse.com/article/61466-artificial-intelligence-ai-utilizing-deep-learning-techniques-to-enhance-adas/, 09 2023.

[2] Artem Oppermann. What Is the V-Model in Software Development? https://builtin.com/software-engineering-perspectives/v-model, 04 2023.

[3] Marie Poenisch. Die Testpyramide. https://www.openknowledge.de/blog/die-testpyramide, 08 2022.

[4] Ham Vocke. The Practical Test Pyramid. https://martinfowler.com/articles/practical-test-pyramid.html, 02 2018.

Erweiterung interner Kurven Berechnungs-Software um eine 3D Simulations-Visualisierung

Andre Glunde **Thomas Rothermel**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Eisenmann GmBH, Böblingen

Einleitung

Stellen Sie sich eine Autofabrik vor: Ein Fördersystem taucht eine Karosserie präzise in ein Lackbad – ein Fehler in der Tauchkurve, wie zu schnelles Auftauchen, ohne dass die Flüssigkeit abfließen kann, verbiegt die Motorhaube. Tauchlackierung ist ein Herzstück der Automobilproduktion, wo Tauchkurven perfekt geplant werden müssen, um Qualität zu sichern [1]. Diese Arbeit erweitert die Avalonia-Anwendung VITRA, die Kurven in 2D-Diagrammen vergleicht, um eine 3D-Visualisierung, die Abweichungen wie Rotationen oder Kollisionen sichtbar macht. Im Fokus stehen Spline-Interpolation für glatte Kurven und Überschleifradien für abgerundete Bahnen, wobei das Ziel ist, Abweichungen schnell zu erkennen und den Mischbetrieb effizienter zu gestalten. Die 3D-Szene zeigt räumliche Details, die in 2D verborgen bleiben, und unterstützt so präzisere Anpassungen. Der Artikel beleuchtet Motivation, Umsetzung, Ergebnisse und Perspektiven, mit Fokus auf die Integration in bestehende Systeme und die Optimierung der Produktion.

Motivation

Tauchkurven richtig zu planen, spart Zeit und Geld. Im Mischbetrieb, wenn alte und neue Fördersysteme parallel laufen, verursachen unterschiedliche Kurven Probleme, z. B. stärkere y-Rotation oder schnellere Beschleunigung. VITRA zeigt Kurven in 2D-Diagrammen für Position (x, z), Rotation (y-Achse), Geschwindigkeit und Beschleunigung. Nutzer prüfen Unterschiede, z. B. 5 Grad y-Rotation oder 0,2 m/s² Beschleunigung, doch räumliche Details bleiben schwer erkennbar. 1 zeigt die 3D-Szene, die solche Abweichungen sichtbar machen soll. Diese Visualisierung beschleunigt die Analyse im Mischbetrieb, wo schnelle Anpassungen entscheidend sind. Optimierte Kurven können die Produktionszeit um bis zu 5 % kürzen [1]. Die 3D-Ansicht ermöglicht es, kritische Punkte wie Kollisionen oder ungünstige Rotationen sofort zu erkennen, was die Fehlersuche und Kurvenoptimierung deutlich vereinfacht.

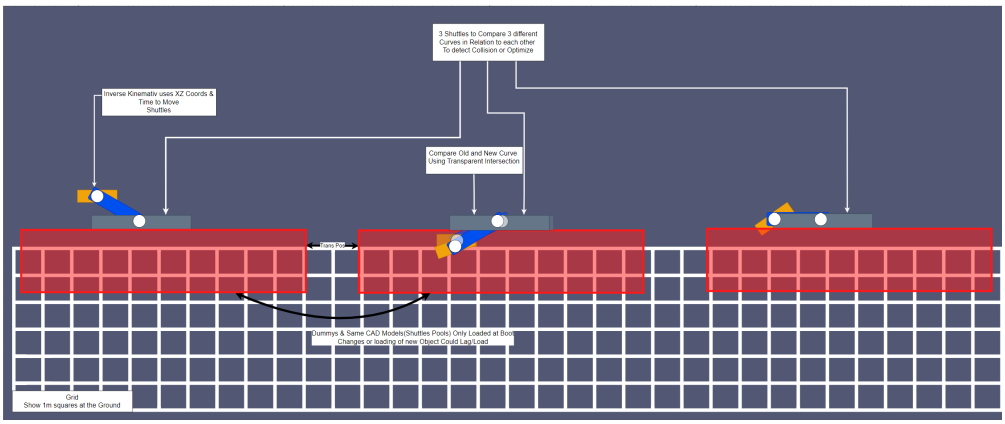


Abb. 1: Schematische Darstellung der 3D-Szene [3]

Zielsetzung

Die Arbeit erweitert VITRA um eine 3D-Visualisierung. Tauchkurven werden in drei Modi verglichen: (1) ein Fördersystem allein, (2) zwei Systeme überlagert mit Transparenz, (3) drei Systeme hintereinander. Kollisionen, z. B. ein Förderarm nahe am Lackbad,

sollen erkannt werden. Modulare CAD-Modelle unterstützen verschiedene Fördersysteme und Karosserien. Die Visualisierung muss in VITRA laufen, ohne 2D-Funktionen zu stören. Im Fokus stehen Spline-Interpolation (Interpolation zwischen Stützpunkten) und Überschleifradien (abgerundete Bahnen). 2 zeigt deren Unterschiede. Ziel ist, Abweichungen leicht erkennbar zu machen und den Mischbetrieb zu optimieren. Die 3D-Szene soll Nutzern helfen, komplexe Bewegungen intuitiv zu verstehen, um Anpassungen schneller und präziser durchzuführen.

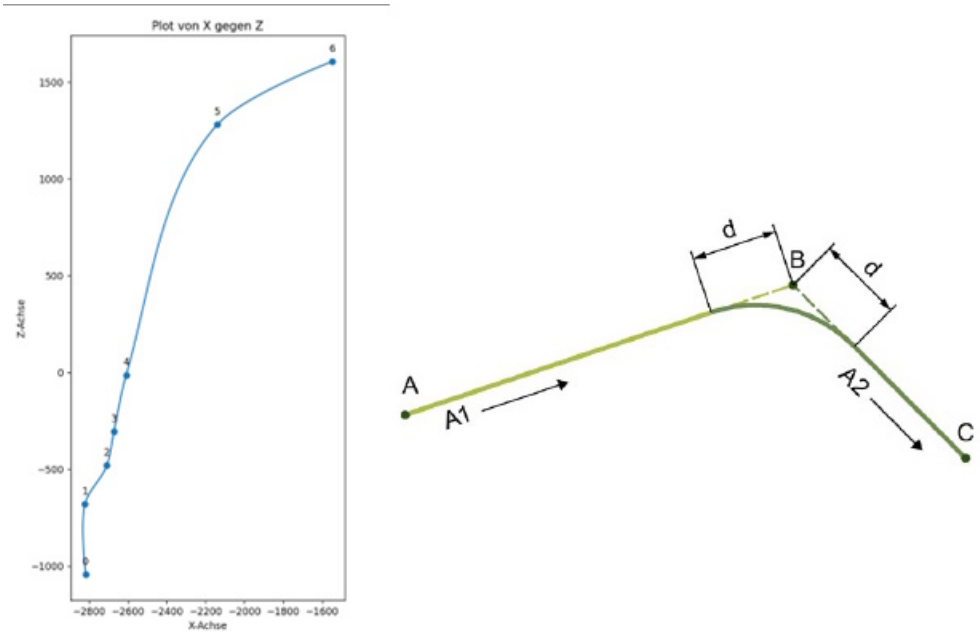


Abb. 2: Kombinierte Darstellung von Spline-Interpolation (intern, Interpolierte Stützpunktverbindungen) und Überschleifradien (extern, abgerundete Bahnübergänge) [3]

Vorgehensweise

Die Anforderungen umfassten das Darstellen mehrerer Fördersysteme, das Erkennen von Kollisionen, die Nutzung modularer Modelle und die Sicherstellung der Kompatibilität mit VITRA. Ein Framework-Vergleich prüfte Unity WebGL, Helix Toolkit, OpenTK und eine robotikspezifische Software. Kriterien waren Implementierungsaufwand, inverse Kinematik, Avalonia-Kompatibilität und Langzeitunterstützung. 3 zeigt die Entscheidungsmatrix, die OpenTK ausweist. OpenTK ist Open-Source, nutzt OpenGL und bietet eine C#-Schnittstelle [5]. Es erstellt eine 3D-Szene und lädt CAD-Modelle via Open Asset Import Library (.obj, .stl). Inverse Kinematik sorgt für realistische Bewegungen, z. B. präzise Kurvenfolge eines Förderarms. Eine dynamische Kamera ermöglicht es Nutzern, die Szene frei zu erkunden, etwa Rotationen von oben oder enge Kurven aus der Nähe zu betrachten. VITRA nutzt das MVVM-Pattern, um Daten und UI sauber zu trennen, sowie Clean Architecture nach Robert C. Martin, um den Code übersichtlich zu gestalten [4]. Dadurch sind Erweiterungen, z. B. für neue Fördersysteme, einfach umsetzbar. Die Visualisierung läuft in einem separaten Fenster, kommuniziert via API mit VITRA und synchronisiert sich mit den 2D-Diagrammen. Nutzer wählen karosserieabhängige Kurven, die in 2D und 3D dargestellt werden, wobei 2D-Diagramme z. B. die x-Position über die Zeit zeigen und die 3D-Szene die Bewegung räumlich darstellt. Die API sorgt für nahtlose Datenübergabe, was die Bedienung intuitiver macht.

		Easy Rob	Unity WebGL	OpenTK	Helix
Kriterium	Gewichtung aus PV	Kriterien	bewertung		
Implementierungsaufwand	10%	3	2	4	4
Inversen Kinematik	8%	3	3	1	1
Overhead	13%	3	1	4	4
Laden von CAD	5%	4	4	1	1
Avalonia-Komplexität	12%	3	3	3	0
Aktuelles Wissen	10%	3	2	4	4
Kollisions Erkennung	7%	4	4	1	1
Änderungen zur Laufzeit	18%	4	3	4	4
Kamera Kontroll	4%	4	4	1	3
Langzeitunterstützung	3%	1	4	4	4
Plattformübergreifend	0%	1	2	3	1
Lizenzgebühr	1%	1	2	4	4
Unabhängigkeit	14%	1	3	4	4
Ergebniss		3,17	2,91	3,39	3,13

Abb. 3: Entscheidungsmatrix für das 3D-Artifact, basierend auf einem paarweisen Vergleich [3]

Ergebnisse

Die 3D-Visualisierung ist in Arbeit, zeigt aber vielversprechende Tests. OpenTK stellt Fördersysteme in drei Modi dar: (1) einzeln, (2) überlagert, (3) hintereinander für Auffahrverhalten. Die dynamische Kamera zeigt Kollisionen aus verschiedenen Winkeln. CAD-Modelle laden flexibel. Optimierungen wie Culling sind geplant. Die API-Anbindung an VITRA ist ebenfalls in Arbeit. Da sehr akribisch mit C4 geplant wurde, ist die Erstellung dieser Bauteile wie das Folgen einer Anleitung. 4 zeigt das C4-Diagramm, das die Architektur verdeutlicht [2]. Es ermöglichte Diskussionen ohne tiefes Code-Verständnis, wobei klar wurde, dass 3D-Objekte und Logik stärker entkoppelt werden mussten, um Clean Architecture einzuhalten [4]. Die Tests zeigen, dass die 3D-Szene komplexe Szenarien wie enge Kurven oder kritische Rotationen klar visualisiert.

Diskussion

Die 3D-Visualisierung übertrifft 2D bei räumlichen Unterschieden: In 2D sieht man, dass die y-Rotation um 5 Grad abweicht, in 3D, ob die Karosserie das Lackbad streift. Die Integration von OpenTK in VITRA war knifflig, da OpenTK OpenGL nutzt, Avalonia aber ein eigenes Rendering. Eine OpenGL-Control löst das, doch Maus-Inputs mussten geparst werden, um flüssige Interaktionen zu ermöglichen. MVVM und Clean Architecture halten den Code sauber [4]. Das C4-Diagramm (Bild 4) zeigt dies klar und erleichterte die Strukturierung [2]. Gegen Unity WebGL ist OpenTK günstiger, bietet aber weniger Features. Zudem muss die 3D-Szene von Grund auf aufgebaut werden, also Shader für Shading aufsetzen und die Projektionsmatrix für die Umwandlung der 3D-Szene zu einer 2D-Ebene. Dennoch bietet OpenTK die nötige Flexibilität für präzise Anpassungen und zukünftige Erweiterungen.

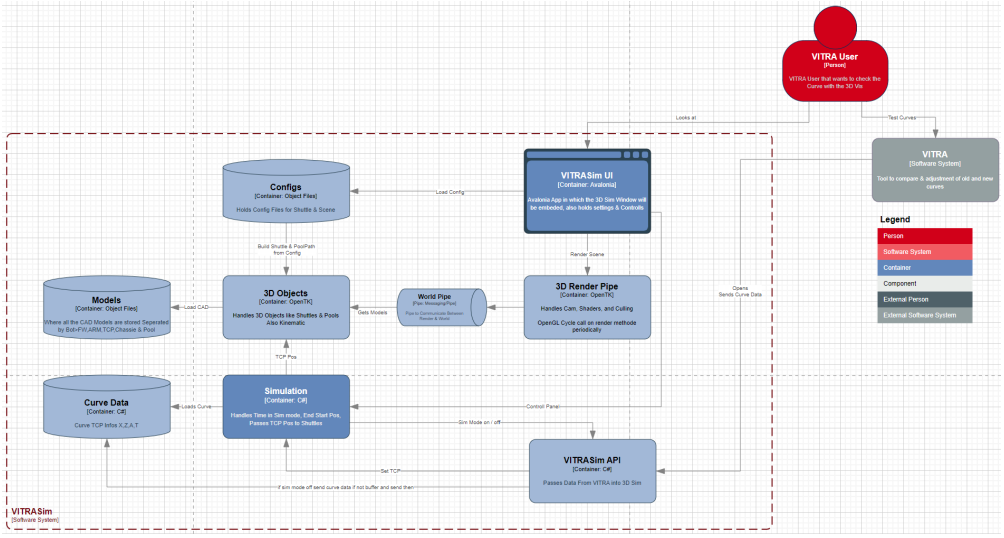


Abb. 4: C4-Diagramm der Software-Architektur, das die 3D-Integration in VITRA zeigt [3]

Ausblick

Nächste Schritte sind die Unterstützung weiterer Tauchlackierungssysteme, um VASIM flexibler zu machen. Performance-Verbesserungen wie Occlusion Culling sollen komplexe Szenen flüssiger machen. Neue Analysewerkzeuge in VITRA, z. B. automatische Warnungen, könnten die Kurvenplanung verbessern. 3D-Visualisierungen haben Potenzial, Fördersystem Planung nachhaltig zu optimieren. Langfristig könnten KI-gestützte Analysen Abweichungen automatisch vorschlagen, um die Effizienz weiter zu steigern.

Literatur und Abbildungen

[1] T. Brock, M. Groteklaes, and P. Mischke. *Lehrbuch der Lacktechnologie.* Vincentz Network, 2000.

[2] S. Brown. The C4 Model for visualising software architecture. https://c4model.com/, 2025.

[3] Eigene Darstellung.

[4] Robert C. Martin. *Clean Architecture: A Craftsman's Guide to Software Structure and Design.* Prentice Hall, 2017.

[5] D. Wolff. *OpenGL 4 Shading Language Cookbook*. Packt Publishing, 2018.

Reporting für Brennstoffzellensysteme: Eine individualisierbare Webapp für die Intralogistik

Max Goehner **Astrid Beck**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Globe Fuel Cell Systems GmbH, Stuttgart

Einleitung

GLOBE Fuel Cell Systems GmbH ist ein deutsches Startup mit Sitz in Stuttgart, das 2020 aus der Brennstoffzellenforschung und dem Innovationsbereich der Mercedes-Benz AG hervorging. Seitdem entwickelt und produziert das Unternehmen Brennstoffzellensysteme für industrielle Anwendungen.
Der Fokus liegt dabei insbesondere auf der Intralogistik, also auf den innerbetrieblichen Material- und Warenflüssen wie Transport, Lagerung und Kommissionierung innerhalb von Logistikzentren oder Produktionsstätten [4]. Typische Anwendungsbereiche sind Flurförderzeuge wie Gabelstapler.
Das Kernprodukt von GLOBE ist der GLOBE XLP80 (vgl. Abb. 1), ein hybrides Brennstoffzellenaggregat, das speziell für den Einsatz in solchen Flurförderzeugen konzipiert wurde.

Abb. 1: Der GLOBE XLP80 – ein hybrides Brennstoffzellensystem für den industriellen Einsatz. [5]

Der XLP80 nutzt einen Brennstoffzellenstack, in dem Wasserstoff (H2) und Sauerstoff (O2) in einer elektrochemischen Reaktion zu Wasser reagieren. Dabei wird elektrische Energie erzeugt, ohne dass Emissionen entstehen. Eine integrierte Lithium-Ionen-Batterie dient als Energiespeicher zur Abdeckung von Lastspitzen, zur Stabilisierung der Stromversorgung und zur Zwischenspeicherung überschüssiger Energie. [5]

Brennstoffzellen in der Intralogistik

Die Nutzung von Brennstoffzellensystemen in der Intralogistik hat in den letzten Jahren an Bedeutung gewonnen, da sie eine nachhaltige und effiziente Alternative zu herkömmlichen Energiesystemen wie Verbrennungsmotoren oder batterieelektrischen Lösungen darstellen. Im Vergleich zu konventionellen Technologien bieten Brennstoffzellen, insbesondere Protonenaustauschmembran-Brennstoffzellen (PEMFC), zahlreiche Vorteile, die sie für die Intralogistik besonders geeignet machen:

- **Schnelles Betanken und hohe Verfügbarkeit:** Wasserstofftanks lassen sich in wenigen Minuten auffüllen, was kurze Standzeiten ermöglicht und den kontinuierlichen Einsatz – auch im Mehrschichtbetrieb – unterstützt.
- **Keine Emissionen:** Bei regenerativ erzeugtem Wasserstoff entstehen weder CO2 noch Schadstoffe, was die Luftqualität in geschlossenen Lagerbereichen verbessert, die Nachhaltigkeit fördert und den Einsatz in Innen- und Außenbereichen, einschließlich Kühlräumen, ermöglicht.
- **Hohe Zuverlässigkeit und Langlebigkeit:** Brennstoffzellen sind robust, wartungsarm und bieten eine konstante Leistungsabgabe bis zur Entleerung des Wasserstofftanks, was sie für den intensiven Mehrschichtbetrieb ideal macht. [2]

Reporting

Im laufenden Betrieb des XLP80 fallen eine Vielzahl technischer Betriebsdaten an – etwa zur Systemlaufzeit oder zum Wasserstoffverbrauch. Diese Betriebsdaten werden kontinuierlich durch Sensoren ausgelesen, gesammelt und zur weiteren Verarbeitung bereitgestellt.

In diesem Kontext rückt das Thema Reporting in den Fokus: Reporting bezeichnet die strukturierte Erfassung, Aufbereitung und Darstellung von Nutzungs- und Leistungsdaten und stellt eine wichtige Grundlage für betriebliche Optimierung, Wartung und strategische Entscheidungen dar. [3]

Zielsetzung

Ziel dieser Arbeit ist die Entwicklung einer prototypischen Reportinglösung, die es Kunden von GLOBE ermöglicht, die im Betrieb des GLOBE XLP80 erfassten Daten effizient auszuwerten und individuelle Reports zu erstellen.

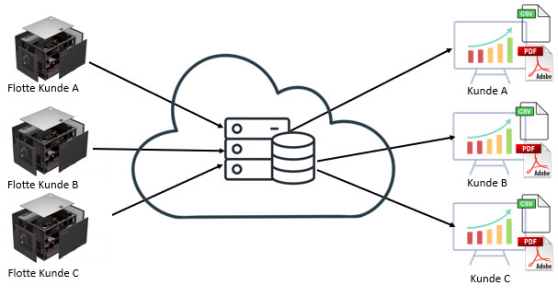


Abb. 2: Betriebsdaten der XLP80-Flotten werden in der Cloud gesammelt und als Reports bereitgestellt [1]

Ein schematischer Überblick über den Datenfluss – von den eingesetzten XLP80-Systemen über die zentrale Datenbank in der Cloud bis hin zum Client ist in Abbildung 2 dargestellt.

Anforderungen

Die zentralen Anforderungen an die Reportinglösung wurden in der Anforderungserhebung gemeinsam mit relevanten Stakeholdern bestimmt und umfassen:

- **Umsetzung als React Web App:** Die Lösung wird als Web-Anwendung auf Basis des React-Frameworks entwickelt, um eine moderne, reaktionsschnelle und plattformunabhängige Benutzeroberfläche zu gewährleisten, die über Standard-Browser zugänglich ist.
- **Hoher Grad an Individualisierbarkeit:** Nutzer können Reports und Dashboards an ihre spezifischen Bedürfnisse anpassen.
- **Exportfunktionen:** Berichte können in die Formate PDF, XLS und CSV exportiert werden, um die Weiterverarbeitung und Integration in bestehende Kundensysteme zu unterstützen.
- **Benutzerfreundlichkeit:** Die Bedienoberfläche ist simpel gestaltet, sodass auch Nutzer ohne IT-Hintergrund die Lösung problemlos nutzen können.

Ansatz

Ein zentrales Feature der Reportinglösung ist das konfigurierbare Dashboard, umgesetzt über ein React Grid (vgl. Abbildung 3). Nutzer können hiermit individuelle Reports erstellen, speichern und jederzeit wiederverwenden. Widgets, wie Datenkacheln oder Diagramme, lassen sich hinzufügen, in der Größe anpassen, per Drag-and-Drop verschieben und exportieren.

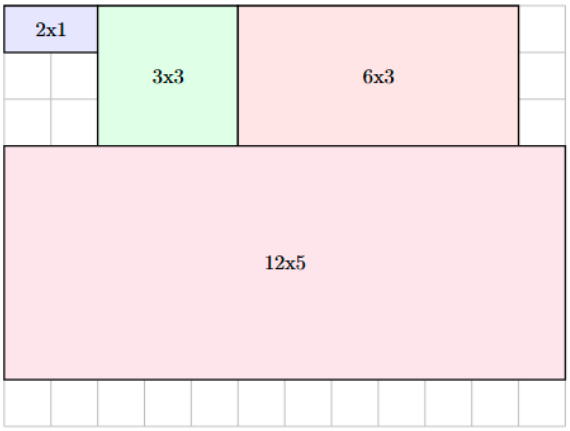


Abb. 3: React Grid mit flexibel konfigurierbaren und verschiebbaren Widgets. [1]

Im Rahmen der Arbeit werden zur Visualisierung der Daten JavaScript-Bibliotheken wie Chart.js und Recharts beleuchtet, die interaktive Diagramme für Kennzahlen zu ermöglichen.
Ein weiteres Feature der Reportinglösung ist ein KI-gestützter Chatbot, der über natürliche Sprachverarbeitung (NLP) gesteuert wird. Dieser ermöglicht intuitive Interaktionen, indem er Anfragen in natürlicher Sprache verarbeitet, wie etwa „Zeige den Wasserstoffverbrauch der gesamten Flotte 'X' des letzten Quartals“ oder „Füge die Betriebsdaten von Einheit 'Y' zu Report 'Z' hinzu“.

Ausblick

Zum Zeitpunkt der Erstellung dieses Artikels befindet sich die Reportinglösung noch in der aktiven Entwicklungs- und Implementierungsphase. Der Fokus liegt aktuell auf der Umsetzung zentraler Kernfunktionen wie der Dashboard-Konfiguration und dem Chatbot. Parallel werden erste Tests durchgeführt, um die Stabilität und Funktionalität der Anwendung sicherzustellen.

Für die nächste Projektphase ist geplant, die Webanwendung in eine produktive Umgebung zu überführen sowie eine praxisnahe Evaluierung der Lösung unter realen Einsatzbedingungen durchzuführen. Insbesondere durch direkte Tests mit Anwendern in der Produktion sollen konkrete Rückmeldungen zur Benutzerfreundlichkeit und zum Funktionsumfang gesammelt werden. Diese Erkenntnisse sollen anschließend in eine iterative Weiterentwicklung der Anwendung einfließen. Auch der Einsatz weiterer KI-gestützter Assistenzfunktionen bietet Potenzial für zukünftige Erweiterungen.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] W.A. Günthner and R. Micheli. H2IntraDrive - Einsatz einer wasserstoffbetriebenen Flurförderzeugflotte unter Produktionsbedingungen. https://www.now-gmbh.de/wp-content/uploads/2020/09/forschungsbericht_h2intradrive_03bs112b.pdf, 2020.

[3] Dietmar Schön. *Planung und Reporting im BI-gestützten Controlling*. Springer Gabler, 2022.

[4] J. Sebulke. Grundlagen der Fördertechnik und der Intralogistik. In *Handbuch Maschinenbau*. Springer Vieweg, 2021.

[5] Liam Sherry. Globe Makes Megawatts From Hydrogen With Simulation. https://www.ansys.com/de-de/blog/globe-makes-megawatts-from-hydrogen-with-simulation, 2023.

Vergleich von Bedien- und Sicherheitskonzepten bei smarten Türschlosssystemen am Beispiel von Homematic IP und Bosch Smart Home

Jonathan Grau **Martin Mink**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Ziel

Bei der Zielsetzung liegt der Hauptfokus auf drei aufeinander aufbauenden Bereichen, die alle das Ziel haben, die Risiken für die Nutzer von Smarthomes zu minimieren:
1. Sicherheitslücken finden: Untersuchung der einzelnen Komponenten und des Gesamtsystems mit dem Augenmerk auf das Auffinden von Sicherheitslücken. Dabei werden sowohl die Hardware als auch die Software betrachtet.
2. Sicherheitslücken analysieren: Die gefundenen Schwachstellen werden mithilfe verschiedener Methodiken analysiert und auf potenzielle Risiken geprüft.
3. Entwicklung von Lösungsideen für Sicherheitslücken: Beschreibung von Verbesserungsvorschlägen für die identifizierten Schwachstellen, um die Risiken eines Angriffs zu minimieren.
Durch diese Arbeit sollen auch Nutzerinnen und Nutzer ohne viel Vorwissen die neuesten Technologien ohne große Bedenken verwenden können. Die Analyse und Optimierung der Sicherheit von Türschlössern im Smarthome stellt dabei nur einen kleinen Teil eines umfassenden Gesamtsystems dar. Die Abbildung 1 veranschaulicht die Montage eines der verwendeten Türschlösser an einem eigens dafür errichteten Versuchsaufbau. Durch diese Testumgebung konnte die einwandfreie Funktion des Schlosses unter realitätsnahen Bedingungen überprüft und sichergestellt werden. Zum Einsatz kamen dabei neben dem Türschloss selbst auch eine zentrale Steuereinheit sowie – im Fall des Homematic-Systems – ein zusätzliches Keypad. Diese Kombination ermöglichte nicht nur die Steuerung und Überwachung des Schlosses, sondern auch eine praxisnahe Simulation alltäglicher Nutzungsszenarien.

Abb. 1: Versuchsuafbau eines Türschlosses [1]

Problembehandlung

Mit der zunehmenden Verbreitung von Smart-Home-Systemen wächst auch die Bedeutung der damit verbundenen Sicherheitsaspekte. Intelligente Türschlösser, Überwachungskameras sowie Licht- und Heizungssteuerungssysteme bieten hohen Komfort und Effizienz – sie eröffnen jedoch zugleich neue Angriffsflächen für Cyberkriminelle. Besonders kritisch ist die Tatsache, dass viele Nutzerinnen und Nutzer keine tiefgreifenden technischen Kenntnisse besitzen, um Sicherheitsrisiken korrekt einschätzen oder absichern zu können. Schwachstellen wie unverschlüsselte Datenübertragung, mangelnde Authentifizierungsmechanismen oder fehlerhafte Firmware-Updates ermöglichen potenziellen Angreifern

das Auslesen sensibler Informationen oder gar die vollständige Kontrolle über sicherheitsrelevante Geräte, wie etwa Tür- und Fensterkontakte. Dabei geht es nicht nur um den digitalen, sondern auch um den physischen Schutz: Ein manipuliertes Smart Lock kann unmittelbare Auswirkungen auf die Sicherheit von Wohnräumen haben. Die Herausforderung liegt also darin, smarte Systeme so zu gestalten, dass sie auch ohne tiefes technisches Verständnis sicher betrieben werden können – und gleichzeitig auf dem aktuellen Stand der IT-Sicherheitstechnologie bleiben.

Forschung

Im Rahmen dieses Projekts wird untersucht, wie sicher Smart-Home-Systeme tatsächlich sind – mit einem besonderen Fokus auf die Kommunikation zwischen den zentralen Steuereinheiten und smarten Türschlössern. Dabei liegt das Hauptaugenmerk auf der Analyse der Funksignale, die zwischen den einzelnen Komponenten ausgetauscht werden. Ziel ist es, potenzielle Schwachstellen in der Datenübertragung aufzudecken. So wurde beispielsweise geprüft, ob sogenannte Replay-Angriffe möglich sind – also Angriffe, bei denen aufgezeichnete Funksignale erneut gesendet werden, um eine unbefugte Aktion auszulösen. Darüber hinaus wurde analysiert, ob sicherheitsrelevante Informationen, wie Zugangscodes für Keypads, aus dem Datenverkehr extrahiert werden können. Ein weiterer Aspekt der Untersuchung ist die Frage, ob sich aus dem Kommunikationsverhalten eine übergeordnete Struktur oder ein Muster ableiten lässt. Eine solche Erkenntnis könnte es Angreifern ermöglichen, eigene – manipulierte – Datenpakete an die Zentrale oder das Türschloss zu senden und so gezielt in das System einzudringen. Langfristig soll die Analyse dazu beitragen, Sicherheitslücken frühzeitig zu erkennen und Lösungen zu entwickeln, die eine sichere Nutzung von Smart-Home-Technologien gewährleisten – selbst für technisch weniger versierte Anwenderinnen und Anwender.

#Ausblick Die Kommunikation zwischen Türschloss und Zentrale ist jedoch nicht der einzige potenzielle Angriffspunkt in einem Smart-Home-System. Auch die Verbindung zwischen der Smartphone-App und der Zentrale stellt ein kritisches Element dar, das umfassend betrachtet werden muss. Angreifer könnten beispielsweise versuchen, sich in diese Verbindung einzuklinken, um Steuerbefehle abzufangen, zu manipulieren oder eigene Kommandos einzuschleusen. Besonders bei unzureichend gesicherten Verbindungen – etwa ohne durchgängige Verschlüsselung oder mit schwacher Authentifizierung – steigt das Risiko erheblich. Neben digitalen Schwachstellen dürfen auch physische beziehungsweise mechanische Angriffspunkte nicht außer Acht gelassen werden. Intelligente Türschlösser basieren trotz ihrer smarten Funktionen nach wie vor auf mechanischen Komponenten, die unter Umständen anfällig für klassische Einbruchsmethoden wie Lockpicking oder gewaltsames Aufbrechen sind. Auch die Montageart und das verwendete Material spielen hierbei eine entscheidende Rolle für die tatsächliche Sicherheit. Ein ganzheitlicher Sicherheitsansatz muss daher sowohl die digitale Kommunikation als auch die physische Beschaffenheit der eingesetzten Komponenten berücksichtigen. Nur durch die Kombination aus robuster Hardware und sicherer Software kann ein effektiver Schutz vor unbefugtem Zugriff gewährleistet werden.

Literatur und Abbildungen

[1] Eigene Darstellung.

Moderne Dashboards und zielgerichtete Visualisierungen – Welche Visualisierungsmethoden fördern bessere Geschäftsentscheidungen?

Melike Guendogan **Dirk Hesse**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

In einer zunehmend datengetriebenen Wirtschaft sind Unternehmen mehr denn je darauf angewiesen, fundierte Entscheidungen auf der Grundlage zuverlässiger Informationen zu treffen. Business Intelligence (BI) hat sich in diesem Zusammenhang als entscheidender Erfolgsfaktor etabliert. Sie ermöglicht es Organisationen, große Datenmengen systematisch zu sammeln, zu analysieren und in nutzbares Wissen zu transformieren. Ein zentrales Element dieser BI-Systeme sind moderne Dashboards, die komplexe Daten visuell aufbereiten und dadurch Entscheidungsprozesse deutlich beschleunigen und verbessern können. Moderne Dashboards dienen nicht mehr nur als statische Berichte, sondern als interaktive Werkzeuge, die Benutzer*innen eine dynamische Analyse in Echtzeit ermöglichen. Sie tragen maßgeblich zur Transparenz und Effizienz in der Informationsverarbeitung bei, indem sie relevante Kennzahlen adressatengerecht visualisieren und somit datenbasierte Entscheidungen unterstützen.
Im Rahmen dieser Abschlussarbeit werden die zentralen Komponenten von Business-Intelligence-Systemen mit Fokus auf die Rolle moderner Dashboards untersucht. Nach einer begrifflichen und konzeptionellen Einordnung von BI wird dargestellt, wie Kernelemente wie Datenquellen, ETL-Prozesse, Data-Warehouse-Strukturen, OLAP-Analysen sowie die visuelle Aufbereitung durch Dashboards ineinandergreifen. Besonderes Augenmerk liegt auf der Funktion von Dashboards als Schnittstelle zwischen Analyse und Entscheidung. Ziel ist es, ein theoretisch fundiertes Verständnis ihrer Einbettung in BI-Systeme zu entwickeln und den Mehrwert zielgerichteter Visualisierungen anhand zentraler Prinzipien und Gestaltungsanforderungen herauszuarbeiten.
Dieser Artikel gibt einen Überblick über die inhaltliche Ausrichtung und die theoretischen Grundlagen der Arbeit.

Definition Business Intelligence

Business Intelligence (BI) bezeichnet die technologiegestützte Erfassung, Aufbereitung, Analyse und Visualisierung von Unternehmensdaten mit dem Ziel, fundierte, datenbasierte Entscheidungen zu ermöglichen. Dabei kommen eine Vielzahl an Prozessen, Werkzeugen und Methoden zum Einsatz, um aus großen Datenmengen relevante Erkenntnisse zu gewinnen und diese in verständlicher Form – etwa in Form interaktiver Dashboards oder Berichte – bereitzustellen [3].
BI wird häufig der deskriptiven Datenanalyse zugeordnet, da vor allem aktuelle und historische Unternehmensdaten im Fokus stehen. Zentrale Fragen wie „Was ist passiert?“ oder „Was muss sich ändern?“ stehen im Mittelpunkt. Ziel ist es, die Unternehmensleistung zu bewerten, Verbesserungspotenziale zu identifizieren und Entwicklungen frühzeitig zu erkennen [5].

Komponenten eines BI-Systems - ein kurzer Überblick

Der Datenverarbeitungsprozess in BI-Systemen folgt einer klaren Struktur: Ziel ist es, aus Rohdaten verwertbare, handlungsorientierte Erkenntnisse zu gewinnen. Zentrale Bausteine sind der ETL-Prozess, das Data Warehouse, analytische Auswertungen (z. B. OLAP) und Dashboards als benutzerfreundliche Oberfläche zur Entscheidungsunterstützung [5].
Der ETL-Prozess (Extract, Transform, Load) bildet die technische Grundlage. Zunächst werden Daten aus diversen internen und externen Quellen extrahiert, etwa aus ERP- oder CRM-Systemen. Anschließend erfolgt die Transformation: Die Daten werden bereinigt, vereinheitlicht und für die Analyse vorbereitet. Schließlich werden sie in ein zentrales Data Warehouse geladen, das strukturierte, konsistente Daten langfristig speichert [4].

Ein Data Warehouse zeichnet sich durch vier Hauptmerkmale aus [1]:

- Subjektorientierung – thematische Gliederung der Daten
- Integration – Vereinheitlichung heterogener Datenquellen
- Zeitraumbezug – Speicherung über definierte Zeiträume
- Nicht-Volatilität – einmal gespeicherte Daten bleiben unverändert.

Diese Eigenschaften ermöglichen verlässliche, langfristige Analysen zur strategischen Unternehmenssteuerung. OLAP (Online Analytical Processing) unterstützt die multidimensionale Analyse großer Datenmengen. In sogenannten „Cubes“ lassen sich Kennzahlen aus verschiedenen Perspektiven (z. B. Zeit, Region, Produktgruppe) flexibel analysieren. Im Gegensatz zu OLTP-Systemen, die auf operative Transaktionen fokussiert sind, dient OLAP der strategischen Entscheidungsfindung und baut auf den Daten aus OLTP auf.
Dashboards stellen die Ergebnisse aus BI-Systemen visuell und interaktiv dar. Sie ermöglichen eine schnelle Erfassung relevanter Kennzahlen, fördern datengestützte Entscheidungen und bieten in Echtzeit einen klaren Überblick über komplexe Sachverhalte [6].
Abbildung 1 zeigt den grundlegenden Aufbau eines Business-Intelligence-Systems – beginnend bei der Datenextraktion aus verschiedenen Quellen über den ETL-Prozess und die zentrale Speicherung im Data Warehouse bis hin zur visuellen Aufbereitung durch Dashboards und Reports.

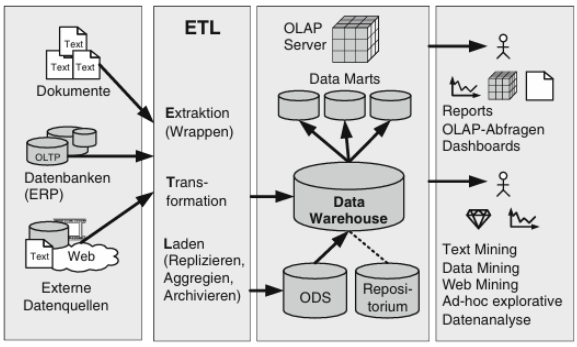


Abb. 1: Prozessübersicht eines BI-Systems [4]

Bedeutung von Dashboards innerhalb von BI-Systemen

Key Performance Indicators (KPIs) sind zentrale Leistungskennzahlen, die den Fortschritt bei der Umsetzung strategischer Unternehmensziele messbar machen. Im Rahmen von Business Intelligence dienen sie der objektiven Bewertung von Prozessen, Abteilungen oder gesamten Organisationseinheiten. Durch definierte Zielgrößen sowie Ober- und Untergrenzen ermöglichen sie eine gezielte Überwachung, Steuerung und Optimierung betrieblicher Abläufe. Ihre visuelle Einbindung in Dashboards verschafft Entscheidungsträgern einen schnellen Überblick über den Status relevanter Geschäftsbereiche.
Ein Dashboard ist eine visuelle Benutzeroberfläche, die entscheidungsrelevante Informationen in komprimierter Form darstellt. Typischerweise werden zentrale Kennzahlen und Inhalte auf einer einzigen Bildschirmseite zusammengeführt, sodass Nutzer*innen die wichtigsten Informationen auf einen Blick erfassen können. Ziel ist es, einen strukturierten Überblick über Leistungsstände oder betriebliche Entwicklungen zu ermöglichen und kritische Sachverhalte durch den Einsatz von intuitiven Visualisierungselementen – wie Diagrammen, Farbcodierungen oder Symbolen – schnell und eindeutig erkennbar zu machen [2].
Zur Unterstützung dieser Interpretation kommen verschiedene Visualisierungselemente zum Einsatz, z. B. Farbsignale (Ampelsymbole), Trendpfeile, Schwellenlinien oder interaktive Diagrammkomponenten (Widgets). Sie erhöhen die Verständlichkeit und erleichtern die frühzeitige Identifikation kritischer Entwicklungen.
Typische Vergleichsdimensionen sind:

- Zeitliche Vergleiche (z. B. Vormonat, Vorjahr)
- Abgleich mit Forecasts oder Budgetwerten
- Zielgrößen und KPIs im Verhältnis zueinander (z. B. Umsatz vs. Marketingausgaben)
- Branchenbenchmarks oder interne Referenzwerte

Durch diese mehrdimensionale Betrachtung erhalten Führungskräfte nicht nur ein aktuelles Lagebild, sondern auch eine fundierte Basis für Ursachenanalysen und strategische Entscheidungen.
Effektive Dashboards verbinden funktionale und gestalterische Prinzipien, um komplexe Informationen klar und verständlich darzustellen. Visualisierungselemente wie Balken- oder Liniendiagramme, Sparklines, Tachometeranzeigen oder farbliche Markierungen machen Abweichungen, Muster und Entwicklungen auf einen Blick erkennbar.
Ein weiterer Erfolgsfaktor ist die Benutzerfreundlichkeit: Dashboards sollten intuitiv bedienbar sein, sodass auch Nutzer ohne technische Vorkenntnisse effizient damit arbeiten können.
Moderne Dashboards konsolidieren Daten aus unterschiedlichen operativen und strategischen Quellen und ermöglichen über Drill-Down- oder Drill-Through-Funktionen eine vertiefte Analyse bis auf Detailebene.

Durch automatisierte, regelmäßige Datenaktualisierungen ist zudem gewährleistet, dass Entscheidungen stets auf aktuellen Informationen basieren [6].

Fazit und Ausblick

Moderne Dashboards sind ein zentraler Bestandteil von Business-Intelligence-Systemen und tragen wesentlich dazu bei, datenbasierte Entscheidungen effizient und fundiert zu treffen. Durch die zielgerichtete Visualisierung relevanter KPIs ermöglichen sie einen schnellen Überblick über betriebliche Entwicklungen und verbessern die Transparenz unternehmerischer Prozesse. Ihre Wirksamkeit hängt jedoch maßgeblich von ihrer Gestaltung, Nutzerfreundlichkeit und Einbettung in die Gesamtarchitektur eines BI-Systems ab.
Zukünftig wird die Bedeutung interaktiver Dashboards weiter zunehmen – insbesondere im Kontext von Self-Service BI und datengetriebenen Organisationen. Technologien wie Künstliche Intelligenz und automatisierte Datenanalyse versprechen zusätzliche Potenziale für die Individualisierung und Automatisierung von Entscheidungsprozessen.

Literatur und Abbildungen

[1] Henning Baars and Hans Georg Kemper. *Business Intelligence & Analytics – Grundlagen und praktische Anwendungen*. Springer Vieweg Wiesbaden, 2021.

[2] Peter Gluchowski, Roland Gabriel, and Carsten Dittmar. *Management Support Systeme und Business Intelligence*. Springer Berlin, Heidelberg, 2008.

[3] . IBM. Was ist Business Intelligence (BI)? https://www.ibm.com/de-de/topics/business-intelligence, 2025.

[4] Roland Müller and Hans Joachim Lenz. *Business Intelligence*. Springer Vieweg Berlin, Heidelberg, 2013.

[5] . SAP. Was ist Business Intelligence (BI)? https://www.sap.com/germany/products/data-cloud/cloud-analytics/what-is-business-intelligence.html, 2025.

[6] Ramesh Sharda, Dursun Delen, and Efraim Turban. *Business Intelligence and Analytics*. Prentice Hall, 2014.

Analyse und Verbesserung von CI/CD-Pipelines: Ein praktischer Ansatz zur Optimierung und Modularisierung bestehender CI-Strukturen

Daniel Hammerschmidt **Dennis Grewe**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma AFI Solutions GmbH, Stuttgart

Einführung

In herkömmlichen Softwareprojekten wurden die Entwicklung (Dev) und Wartung (Ops) der Software voneinander getrennt behandelt. Dies sorgte für langsamere Rückmeldung, und hatte damit auch eine langsamere Anpassung und Verbesserung der Software zur Folge. In moderneren Projekten wird deswegen immer häufiger auf einen Zusammenschluss der beiden Aspekte gesetzt. Dieser wird als DevOps (Development and Operation) bezeichnet, und soll dazu beitragen, die Geschwindigkeit von Entwicklung, Auslieferung und Pflege zu beschleunigen. Ein Schlüsselkonzept ist hierbei die Verwendung von Continuous Integration/Continuous Deployment/Delivery (CI/CD) [2]. Eine Übersicht über alle DevOps-Aspekte ist in Abbildung 1 dargestellt. Aspekte, die zu CI/CD gehören, sind in der Abbildung rot markiert.

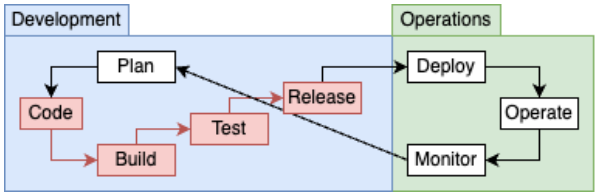


Abb. 1: DevOps-Lebenszyklus mit hervorgehobenen CI/CD-Aspekten [1]

Bei CI/CD handelt es sich um Methoden, welche darauf abzielen Rückmeldungs- und Deployment-Prozesse zu beschleunigen, um eine insgesamt kürzere Releasezeit zu erhalten. Die Prozesse erfolgen dabei automatisiert auf einer dedizierten Infrastruktur. CI bedeutet hierbei, dass Änderungen in der Codebasis in regelmäßigen Zeitabständen in ein geteiltes Quellcodeverzeichnis (Repository) integriert werden. Das Ziel besteht darin, möglichst schnell aus diesen Änderungen eine angepasste und qualitativ hochwertige Softwarekomponente als Build-Artefakt zu erhalten, welche durch die Automatisierung von Test- und Build-Prozessen erzielt wird. CI trägt somit wesentlich zur schnellen Rückmeldung und frühzeitigen Fehlererkennung bei. CD, oft als Akronym für Continuous Deployment oder Continuous Delivery, steht für das automatisierte Deployment von Softwarekomponenten. Die beiden Herangehensweisen unterscheiden sich in dem Grad der Automatisierung. Continuous Delivery sorgt dafür, dass durch CI gebaute Softwarekomponenten bereit für ein Deployment sind, welches jedoch eine manuelle Freigabe benötigt. Continuous Deployment setzt an dieses Prinzip an und automatisiert die Auslieferung der Softwarekomponente an die Infrastruktur des Kunden [4]. Durch ihre inkrementelle Herangehensweise eignen sich diese Methoden für agile Formen des Projektmanagements wie Scrum.

Motivation und Zielsetzung

Wie zuvor erläutert, verfolgt CI den Zweck, Releasezeiten durch Automatisierung zu verkürzen. Diese setzten sich durch die Kombination von Entwicklungs- und Build-Zeiten zusammen. Dementsprechend lohnenswert ist es, neue CI-Strukturen aufzubauen und bestehende Strukturen zu optimieren. Die Automatisierung trägt dazu bei, dass die Rückmeldung ohne manuelle Ausführung eines lokalen Build-Prozesses erhalten werden kann. Eine typisches Beispiel ist hierbei die Arbeit an einer Änderung, die über einen Push zu einem Repository eines Cloud-Anbieters hinzugefügt wird. Eine dort eingerichtete Pipeline prüft die eingebrachte Änderung automatisch. Vorteil hier ist, dass nahezu unterbrechungsfrei weitergearbeitet werden kann. Die Pipeline läuft bei dem Cloud-Dienst und gibt innerhalb von wenigen Minuten eine Rückmeldung. Darüber hinaus müssen keine lokalen Ressourcen aufgewendet werden, welche somit frei für die Verwendung in der Entwicklung sind. Damit wird die Entwicklungszeit verkürzt, da manuelle Tätigkeiten

wegfallen. Die technische Realisierung erfolgt durch eine CI/CD-Pipeline. Hierbei handelt es sich um eine Sammlung von Arbeitsschritten (Jobs), welche automatisiert ausgeführt werden. In Abbildung 2 wird ein exemplarischer Ablauf einer solchen Pipeline dargestellt.

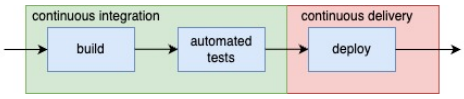


Abb. 2: Typisches CI/CD-Ablaufdiagramm [1]

Neben der Entwicklungszeit ist auch die Dauer, die ein Build-Vorgang benötigt, ein zentraler Aspekt. Durch eine Anpassung der bestehenden CI-Skripte können Vorgänge, wie zu Beispiel Builds, signifikant beschleunigt werden. Hierzu gehören auch strukturelle Themen, wie die Ausführungsreihenfolge von Jobs und die Parallelisierung von untereinander unabhängigen Jobs. Darüber hinaus senkt eine Modularisierung der Pipeline den Entwicklungsaufwand. Insbesondere in größeren Projekten mit vielen Pipelines ist es möglich, dass Jobs identisch sind. Eine Modularisierung würde an dieser Stelle dazu beitragen, den Aufwand für die Anpassung mehrerer Pipelines zu senken. Damit geht eine Reduktion der Entwicklungszeit einher, welche im Kontext der Reduktion von Releasezeiten ebenfalls einen positiven Beitrag leistet.

Analyse bestehender Strukturen und Metriken

Um eine Analyse der bestehenden Pipeline-Struktur durchzuführen, soll zunächst ein Überblick über diese geschaffen werden. Es handelt sich bei der bisher um mehrere Teilprojekte, welche jeweils eine eigene Pipeline besitzen. Der Fokus dieser Pipelines liegt hierbei insbesondere auf der Zusammenführung von neuen Code-Branches (Feature-Branches) in den Hauptbranch (Main-Branch), welche auch als Merge bezeichnet wird. Diese Folgen dabei dem typischen CI/CD-Ablauf aus Abbildung 2. Das bedeutet auch, dass die Pipelines untereinander einen ähnlichen Aufbau haben. So besitzt jede Pipeline die selbe Struktur an logischen Gruppen (Stages). Jede Stage hat dabei einen gleichnamigen Job, welcher die namensgebende Aktivität ausführt. Dabei führen Jobs mit dem selben Namen die selben Aktivitäten durch. Sie sind – je nach Projekt und Programmiersprache – ähnlich aufgebaut und unterscheiden sich oft nur durch angegebene Namen und Verzeichnispfade. Dies hat zur Folge, dass mehrere Pipelineskripte angepasst werden müssen, sollten sich Konfigurationen ändern. Ein Beispiel dafür ist eine Änderung einer URL, welche unter Umständen in mehreren Pipelines angepasst werden müssen. Darüber hinaus sind Build-Vorgänge oft langsam. Das liegt daran, dass jeder Pipeline-Job in einer eigenen Umgebung ausgeführt wird. Der Aufwand, welcher dadurch entsteht, ist nicht unerheblich für die Ausführungsdauer. Hierzu gehört in erster Linie die Initialisierung der Build-Umgebung zu Beginn des Build-Vorgangs. Abhängigkeiten von Fremdkomponenten (Dependencies) der zu bauenden Softwarekomponente müssen bei jedem Build erneut aufgelöst und heruntergeladen werden. Viele Werkzeuge für Software-Builds (Build-Tools), wie Gradle stellen Funktionalitäten bereit, um diese Vorgänge zu beschleunigen oder zu vermeiden. Dies ist aber nur möglich, wenn die Build-Umgebung persistiert wird. Ein weiterer Faktor, welcher die Geschwindigkeit einer CI/CD-Pipeline beeinflusst, ist die sequenzielle Ausführung von Jobs innerhalb einer Stage. Ein Projekt, welches aus mehreren Child-Pipelines besteht, kann durch die sequenzielle Ausführung der Child-Pipelines ausgebremst werden. Die ausschlaggebende Metrik ist in diesem Falle einerseits die Durchlaufzeit der Pipeline, andererseits die Durchlaufzeit der Schritte, welche Potenzial für Verbesserungen bieten. Somit bietet die aktuelle Pipelinestruktur ein deutliches Potenzial, um die beiden zeitlichen Hauptfaktoren zu beschleunigen. Die Entwicklung und Anpassung von Pipelines stellt dabei einen Entwicklungsaufwand dar, welcher durch die Modularisierung mithilfe von vorgefertigten Komponenten reduziert werden kann. Durch den Einsatz von Methoden zur Persistenz der Build-Umgebung lassen sich Build-Vorgänge signifikant beschleunigen. Darüber hinaus kann dadurch die Nutzung der von den Build-Tools bereitgestellten Funktionalitäten zur Build-Beschleunigung ermöglicht werden.

Konzeptionelle Umsetzung und Teilergebnisse

Viele Pipeline-Plattformen wie Jenkins, GitLab CI und GitHub Actions bieten Funktionalitäten zur Modularisierung von Pipelines in Komponenten. Damit ist es möglich, redundante Schritte auszulagern, um dem Don't-repeat-yourself-Prinzip (DRY) zu folgen. Die Analyse der bestehenden Pipeline-Struktur hat ergeben, dass die verwendeten Stages und Jobs marginale Unterschiede besitzen. Damit ist eine Modularisierung möglich. Ein Beispiel hierfür ist die Deployment-Stage, welche sich nur durch verwendete Namen und Dateipfade unterscheidet. Denkbar ist hier die Auslagerung in eine einzelne, zentral verwaltete Komponente, welche in die aktuellen Pipelines eingesetzt werden kann. Um komponentenspezifische Informationen, wie Namen oder Dateipfade zu steuern, können Umgebungsvariablen verwendet werden. Diese sollen in der komponentenspezifischen Pipeline gesetzt, und anschließend in dem Modul verwendet werden.

Build-Zeiten lassen sich durch das Caching der Build-Umgebung beschleunigen [3]. Der Aufwand, welcher durch die Initialisierung der Umgebung stattfindet, lässt sich damit reduzieren, wenn nicht gar vermeiden. Durch das Caching der Build-Umgebung sind alle benötigten Dateien, wie Dependencies, sofort verfügbar. Ein erneutes Herunterladen ist somit nicht nötig und kann vermieden werden. In [3] wird hierfür der Einsatz von eigens angelegten Docker Images beschrieben, welche nach einem initialen Build-Vorgang (Cold Build) zur Verwendung von weiteren Bauvorgängen (Warm Build) verwendet werden. Dei einem Docker Image handelt es sich um einen Bauplan für einen Docker Container, welcher die Basis für die Ausführungsumgebung eines Jobs darstellt. Diese Arbeit verwendet eine abgeschwächte Variante: es sollen nur Verzeichnisse und Dateien gespeichert werden, welche relevant für den Build-Vorgang sind. Dieser Vorgang wird als Dependency Caching bezeichnet. Damit lässt sich die Umgebung schnell wiederherstellen, und der Aufwand für die Erstellung eines Docker Images fällt weg. Die Umsetzung dieser Praxis hat bereits eine Verbesserung um durchschnittlich 70% ergeben. Gemessen wurde hierbei die Differenz aus dem Zeitstempel (Timestamp) vor Beginn und nach Ende des Build-Vorgangs in der Konsole der Job-Umgebung. Die in der Pipeline gebauten Software-Artefakte wurden darüber hinaus erfolgreich auf ihre Integrität geprüft. Hierfür wurden Pipeline-Artefakte mit Artefakten aus lokalen Builds bitweise verglichen. Dabei ergaben sich abgesehen von Timestamps jedoch keine weiteren Unterschiede zwischen den Artefakten. Pipelines, welche aus mehreren Child-Pipelines bestehen, können durch die parallelisierte Ausführung dieser zusätzlich beschleunigt werden, da es sich hier um unabhängige Softwarekomponenten handelt. Eine kombinierte Anwendung dieser Methodik mit Dependency Caching ermöglichte eine Reduktion der Pipeline-Durchlaufzeit um bis zu 65%.

Fazit und Ausblick

Die Umsetzung des Dependency Caching ermöglichte eine signifikante Reduktion der Build-Vorgänge und ermöglicht damit eine insgesamt schnellere Releasezeit. Durch die zusätzliche Verwendung von Parallelisierung können diese Zeiten weiterhin reduziert werden, was ebenso einen positiven Effekt auf die Releasezeiten hat. Die Modularisierung der Pipelines stellt einen wichtigen Schritt Richtung Wartbarkeit dar. Es wird damit einfacher möglich sein, Pipelines für neue Projekte aufzusetzen. Darüber hinaus stellt es ein einfaches Verfahren zur schnellen Anpassung dar. Dies zeigt sich in größeren Projekten, welche aus mehreren Pipelines bestehen. Die Notwendigkeit, viele einzelne Skripte anzupassen, entfällt. Denkbar wäre hier auch ein Streamlining der Pipelinekomponenten in einem abteilungs- und gegebenenfalls unternehmensweiten Kontext. Komponenten könnten damit zentralisiert gespeichert und gewartet werden. Damit wird es den Entwicklerteams ermöglicht, ihre gesamten Kapazitäten auf die Weiterentwicklung ihres Produktes zu leiten. Die Verwendung von Dependency Caching kann darüber hinaus einen ersten Schritt in Richtung deterministische Kompilation (Reproducible Builds) setzen, welche das Ziel hat, bei jedem Build-Vorgang einen bitweise identischen Output zu erzeugen.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Christof Ebert, Gorka Gallardo, Josune Hernantes, and Nicolas Serrano. DevOps. *IEEE Software*, 33:94–100, 2016.

[3] Keheliya Gallaba, John Ewart, Yves Junqueira, and Shane McIntosh. Accelerating Continuous Integration by Caching Environments and Inferring Dependencies. *IEEE Transactions on Software Engineering*, 48:2040–2052, 2022.

[4] Mojtaba Shahin, Muhammad Ali Babar, and Liming Zhu. Continuous Integration, Delivery and Deployment: A Systematic Review on Approaches, Tools, Challenges and Practices. *IEEE Access*, 5:3909–3943, 2017.

Entwicklung eines LLM basierten, interaktiven Wissenssystems zur Analyse von Software-Dokumentation unter Berücksichtigung der Ressourceneffizienz

Marcel Hartmann **Dennis Grewe**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Mensch und Maschine Infrastruktur, Ditzingen

Einführung

Die effiziente Erlangung von technischem Wissen ist entscheidend für die Verbesserung von Arbeitsabläufen in der Softwareentwicklung und im Beratungsbereich. In der Praxis führen unstrukturierte oder schwer zugängliche Informationen oft zu einem Anstieg des Kommunikationsaufwands und Zeitverlust. Das Ziel dieser Arbeit besteht darin, ein Wissenssystem mit künstlicher Intelligenz zu entwickeln, das lokal ausgeführt werden kann und den Schutz sensibler Daten sicherstellt. Mithilfe von großen Sprachmodellen (LLMs), Retrieval-Augmented Generation (RAG) und der Anpassung an spezifische Domänendaten soll eine strukturierte Erschließung von technischen Dokumentationen und Quellcode erreicht werden. Dabei werden die Inhalte miteinander verknüpft, können bei Bedarf auch einzeln betrachtet und über eine einheitliche Benutzeroberfläche zugänglich gemacht werden.. Die Bewertung des Systems basiert auf qualitativem Feedback der Nutzer, um die Eignung und den Mehrwert in der Praxis zu beurteilen.

Problemstellung

Die Entwicklung eines KI-unterstützten Wissenssystems, das lokal betrieben wird, stellt besondere technische Probleme dar. Um die Sicherheit sensibler Daten zu gewährleisten, ist es erforderlich, dass das System keine Kommunikation nach außen gewährleistet und vollständig unabhängig von externen Servern oder Cloud-Services arbeitet. Die beschränkte Verfügbarkeit von lokalen Rechenressourcen stellt gleichzeitig eine bedeutende Schwierigkeit dar. Es ist notwendig, die Steuerung von verschiedenen Large Language Models (LLMs) mit unterschiedlichem Ressourcenbedarf je nach Komplexität der Nutzeranfragen anzupassen, um ein Gleichgewicht zwischen Antwortqualität und Effizienz zu gewährleisten. Ein flexibles Architekturkonzept ist notwendig, um sowohl eine effiziente Nutzung des Modells als auch eine zufriedenstellende Genauigkeit der Antworten zu gewährleisten. Es ist erforderlich, eine benutzerfreundliche Oberfläche zu entwickeln, die den intuitiven Zugriff auf Informationen ermöglicht. Der Bedarf an fortgeschrittenen KI-Technologien und einer durchdachten Systemarchitektur entsteht aufgrund der Anforderungen an Datenschutz, Effizienz und Benutzerfreundlichkeit.

Entwicklung des wissensbasierten Assistenzsystems

Für die Umsetzung des wissensbasierten Assistenzsystems wurde eine strukturierte Architektur entwickelt, die Effizienz, Datenschutz und die Verarbeitung domänenspezifischen Wissens vereint. Die zugrunde liegende Logik basiert auf einem mehrstufigen, modularen System mit einer Kombination aus Retrieval-Augmented Generation (RAG) und spezialisierten, feinjustierten Sprachmodellen.

Wie in Abbildung 1 dargestellt, werden technische Dokumentationen im Markdown-Format eingelesen und zunächst in logisch segmentierte Textabschnitte unterteilt. Diese Abschnitte werden mittels vortrainierter Embedding-Modelle in Vektoren umgewandelt und in einer Vektor-Datenbank abgelegt. Diese Vorverarbeitung erfolgt im Modul Service-Embeddings, das für das Chunking und die Erzeugung semantischer Repräsentationen zuständig ist.

Die Nutzerinteraktion erfolgt über einen Chatbot, der auf einem modernen Frontend (Angular/NodeJS/NGINX) basiert. Eingehende Anfragen werden über REST-Requests an das zentrale Backend-Modul (Service-Backend-Control) weitergeleitet, das mit Python/Django umgesetzt ist und einen RAG-Controller enthält. Dort werden zunächst relevante Vektoren zur Anfrage aus der Datenbank abgefragt, um kontextuelle Informationen zu ermitteln.

Im nächsten Schritt wird die Anfrage durch ein leichtgewichtiges Sprachmodell (Lightweight LLM) analysiert, um den Anfrage-Typ zu klassifizieren. Auf Basis dieser Klassifikation wird entschieden, welches Sprachmodell verwendet wird: entweder ein Fine-Tuned LLM für inhaltlich-thematische Anfragen oder ein spezialisiertes Fine-Tuned Code LLM für technische und programmierbezogene Inhalte (z. B. Quellcode, APIs, Fehlermeldungen).
Diese Architektur gewährleistet eine adaptive Modellauswahl, optimale Ressourcennutzung sowie eine hohe Qualität und Präzision der generierten Antworten – stets abgestimmt auf Struktur und Komplexität der jeweiligen Anfrage.

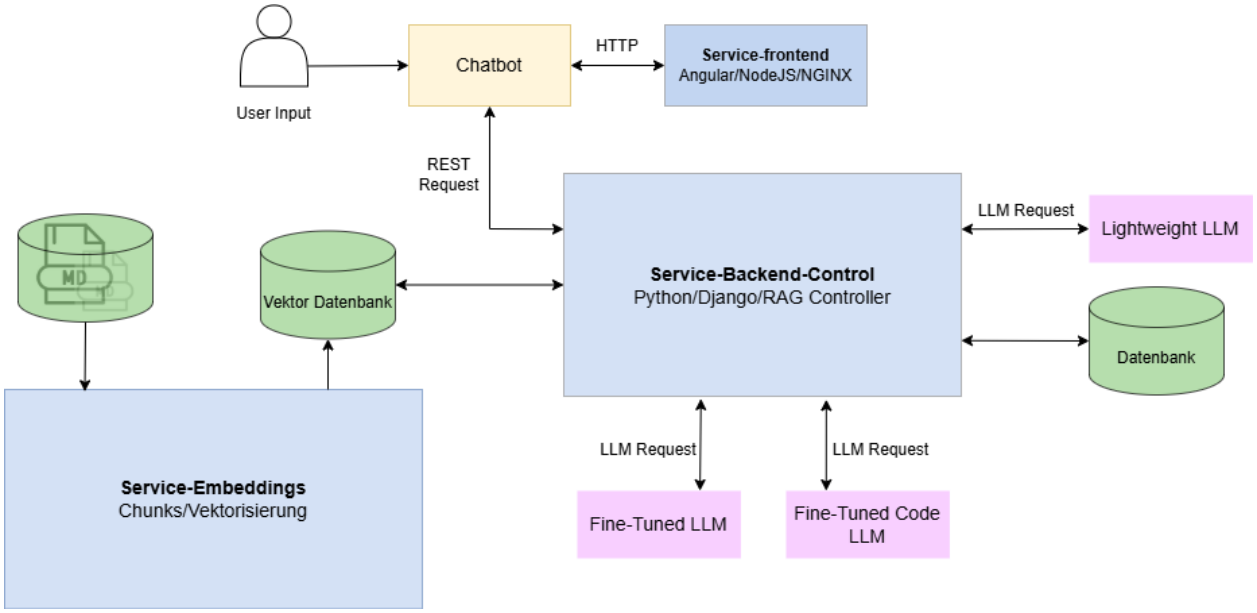


Abb. 1: Architektur des Prototyps [2]

Schwierigkeiten im Verlauf der Systementwicklung

Es gibt verschiedene technische und konzeptionelle Schwierigkeiten bei der Erstellung eines lokal betriebenen Assistenzsystems, das auf Large Language Models basiert. Ein bedeutendes Problem liegt darin, dass herkömmliche Sprachmodelle Schwierigkeiten haben, umfangreiche Mengen an strukturierten Textdaten, vor allem viele Markdown-Dokumente, effektiv zu verarbeiten. Je mehr Inhalte integriert werden, desto wahrscheinlicher ist es, dass das Modell ungenaue oder fiktive Antworten generiert. Diese Erscheinung hat negative Auswirkungen auf die Präzision und Zuverlässigkeit der erstellten Ergebnisse. Ein weiteres Hauptproblem entsteht bei der Bewertung der Musterlösungen. Es ist für Benutzer nicht sofort ersichtlich, ob die Antwort der Sprachmodelle mit den zugrunde liegenden Daten übereinstimmt, da sie keine klaren Verweise auf spezifische Abschnitte der Dokumentation liefern. Um die Ausgaben zu überprüfen, ist es notwendig, eine zusätzliche Infrastruktur einzurichten, die beispielsweise die Möglichkeit bietet, qualitatives Feedback von Benutzern zu erhalten oder manuelle Prüfprozesse zu unterstützen. Eine zusätzliche Schwierigkeit besteht darin, verschiedene Informationsquellen wie technische Handbücher und Programmcode effektiv miteinander zu verknüpfen. Die Struktur und der sprachliche Stil variieren stark zwischen den verschiedenen Arten von Quellen, weshalb spezielle Sprachmodelle für jede Art benötigt werden. Um den Quellcode zu bearbeiten, benötigt man ein Modell, das sich auf Programmierkontexte konzentriert, während für die Auswertung der Dokumentation ein Modell mit umfassendem Verständnis natürlicher Sprache erforderlich ist. Um sicherzustellen, dass das System angemessen genutzt werden kann, sollte es fähig sein, Anfragen zu untersuchen und basierend auf deren Inhalten und Bedürfnissen das passende Sprachmodell auszuwählen. Zusätzlich spielt die effiziente Nutzung der Modelle eine wichtige Rolle. Die Leistungsfähigkeit eines Sprachmodells ist unmittelbar von der Anzahl der Tokens in einer Anfrage abhängig. Um sicherzustellen, dass das System effizient arbeitet, muss es die Anzahl der Tokens in einer Anfrage messen und dann ein passendes Modell auswählen, das sowohl hinsichtlich Größe als auch Leistung optimal zur spezifischen Anfrage passt. Es soll vermieden werden, umfangreiche Modelle außer bei komplexen Fragen zu verwenden, um die verfügbaren Rechenressourcen effizient zu nutzen und eine skalierbare Systemstruktur sicherzustellen.

Ergebnisse und Ausblick

Das Ziel des entwickelten Assistenzsystems besteht darin, genaue Antworten auf Fragen zu liefern, die sowohl die Dokumentation als auch den Quellcode betreffen. Die Struktur des Systems vereint Retrieval-Augmented Generation mit maßgeschneiderten Sprachmodellen und legt besonderen Wert auf die Berücksichtigung von Ressourceneffizienz durch eine flexible Auswahl der Modelle. Je nach der Komplexität und Art der Anfrage wird entweder ein Modell verwendet, das auf die technische Dokumentation zugeschnitten ist, oder ein leistungsstärkeres Modell, das auf Programmierkontexte spezialisiert ist. Aktuell befindet sich das System in einer Phase des Prototyps. Es sind erste Nutzertests geplant, um systematisch die Genauigkeit und Relevanz der generierten Antworten zu überprüfen und mögliche Schwachstellen in der Modellwahl oder den Retrieval-Ergebnissen aufzudecken. Das Ziel dieser Prüfungen besteht darin, die Wirksamkeit des Vorgehens durch empirische Bestätigung zu belegen und den Weg für zukünftige Verbesserungen zu ebnen. Für zukünftige Erweiterungen bestehen verschiedene Perspektiven. Eine Möglichkeit stellt die Integration multimodaler Fähigkeiten dar, etwa durch die Einbindung von Bildgenerierung zur Visualisierung technischer Inhalte. Ein zentrales Hindernis beim praktischen Einsatz von KI-Systemen ist die begrenzte Anbindung an kontextspezifische Datenquellen. Häufig arbeiten Modelle getrennt von den Systemen, in denen wichtige Informationen gespeichert sind, was ihre Anwendbarkeit in komplexeren Szenarien begrenzt.

Das Model Context Protocol (MCP) stellt einen offenen Standard bereit, um externe Datenquellen zu integrieren. MCP stellt eine einheitliche, standardisierte Schnittstelle bereit, um KI-Assistenten sicher und flexibel mit verschiedenen Datenquellen wie Dokumentenablagen, Code-Repositories oder Geschäftsanwendungen zu verknüpfen. MCP ermöglicht eine nachhaltige und skalierbare Integration anstelle der Entwicklung spezifischer Konnektoren für jede Plattform, ähnlich dem Aufkommen von APIs im Internet.

Auf diese Weise kann die Integration von KI-Systemen in vorhandene Informationsumgebungen verbessert und die automatisierte Bearbeitung kontextabhängiger Aufgaben erleichtert werden. Dies führt zu neuen Anforderungen bezüglich Datenschutz und Systemarchitektur, die später berücksichtigt werden müssen [1].

Literatur und Abbildungen

[1] PBC Anthropic. Introducing the Model Context Protocol. https://www.anthropic.com/news/model-context-protocol, 11 2024.

[2] Eigene Darstellung.

Radarbasierte Konturenerkennung durch maschinelles Lernen mit LiDAR-Referenzdaten

Salen Hasanovic **Clemens Klöck**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

Radarsensoren (Radio Detection and Ranging) stellen eine etablierte Technologie zur Erfassung von Umgebungsdaten dar. Mithilfe hochfrequenter Radiowellen können Informationen wie die Entfernung, die Relativgeschwindigkeit sowie der Winkel bestimmt werden, unter dem sich Objekte relativ zum Sensor befinden. Ursprünglich im militärischen Bereich eingesetzt, finden Radarsysteme heute Anwendung in zahlreichen zivilen Branchen. Insbesondere in der Automobilindustrie gelten sie als unverzichtbare Komponente für Fahrerassistenzsysteme wie Querverkehrswarnungen (RCTA), Spurwechselassistenten, Kollisionsvermeidung und Totwinkelerkennung (BSD) [2]. Auch in Bereichen wie Medizin, Raumfahrt und industrieller Prozessautomatisierung gewinnt die Radartechnologie zunehmend an Bedeutung. Diese Entwicklung ist vor allem auf technologische Fortschritte zurückzuführen: Moderne Radarsensoren sind kompakter, kostengünstiger und leistungsfähiger als frühere Generationen – bedingt durch optimierte Hochfrequenzkomponenten, fortschrittliche Gehäusetechnik und effizientere Fertigungsprozesse [2].

Abb. 1: Marktübersicht von Radarsensoren weltweit [2]

Das wirtschaftliche Potenzial dieser Technologie zeigt sich deutlich im prognostizierten Marktwachstum (siehe 1]). Für das Jahr 2024 wurde ein Marktvolumen von 21,03 Milliarden US-Dollar ermittelt; bis 2029 wird ein Anstieg auf 45,42 Milliarden US-Dollar erwartet. Die zunehmende Marktdurchdringung unterstreicht die Rolle von Radarsensoren als Schlüsseltechnologie in modernen Sensorsystemen [2].

Zielsetzung:

Ziel dieser Arbeit ist die Untersuchung des Potenzials von Radarsensoren zur Erfassung von Objektkonturen im Vergleich zu LiDAR-Systemen. Zu diesem Zweck wird ein Testsystem konzipiert, das die Daten beider Sensortechnologien synchron und in Echtzeit erfasst. Die erfassten Radar- und LiDAR-Daten werden in einem gemeinsamen Koordinatensystem visualisiert, um einen direkten visuellen und quantitativen Vergleich zu ermöglichen. Die LiDAR-Daten dienen dabei als Referenz (Ground Truth), anhand derer die Qualität und Genauigkeit der Radarmessungen bewertet werden können. Auf Grundlage dieser Datenbasis wird ein neuronales Netz trainiert, das in der Lage ist, Radar-3D-Punktwolken in LiDAR-3D-Punktwolken zu überführen. Aufgrund der höheren strukturellen Auflösung von LiDAR-Systemen fungieren deren Punktwolken als Zielausgabe für die Rekonstruktion. Voraussetzung für ein erfolgreiches Training ist, dass die vom Radarsensor erzeugten Punktwolken eine ausreichende Genauigkeit aufweisen, um eine realistische Annäherung an LiDAR-Qualität zu ermöglichen. Die Fähigkeit zur Echtzeitverarbeitung stellt dabei einen zentralen Aspekt dar – insbesondere im Hinblick auf potenzielle Anwendungen in dynamischen und sicherheitskritischen Umgebungen.

Analyse des Cascade-Radarsensors:

Aufgrund des Umfangs des Projekts erfolgt die Bearbeitung durch zwei Studierende in paralleler Zu-

sammenarbeit mit klar definierter Aufgabenverteilung. Der hier dokumentierte Teil befasst sich mit der Analyse des hochauflösenden Cascade-Radarsensors von Texas Instruments, bestehend aus den Modulen MMWCAS-RF-EVM und MMWCAS-DSP-EVM. Im Mittelpunkt steht die Untersuchung der vollständigen Signalverarbeitungskette – von der Erfassung des Antennensignals bis zur Generierung der finalen 3D-Punktwolken. Zusätzlich werden wesentliche Leistungsmerkmale wie Reichweite, Winkelauflösung, Entfernungsauflösung sowie die Echtzeitfähigkeit des Systems analysiert. Ergänzend dazu erfolgt die Dokumentation der Kalibrierverfahren sowie der spezifischen Anforderungen und Besonderheiten bei der Inbetriebnahme.

FMCW-Radar

Ein Radarsystem nutzt hochfrequente elektromagnetische Wellen, um Objekte zu detektieren und deren Entfernung, Relativgeschwindigkeit sowie Richtung (Azimut/Elevation) zu bestimmen. Es besteht aus einem Sender, der das Signal aussendet, und einem Empfänger, der die vom Objekt reflektierten Signale aufnimmt. Die Intensität dieser Echosignale hängt von den Material- und Formeigenschaften des reflektierenden Objekts ab und wird durch die sogenannte Radarquerschnittsfläche (Radar Cross Section, RCS) beschrieben. Aus den empfangenen Signalen lassen sich Rückschlüsse auf Position und Bewegung der Objekte ziehen. Eine in der Automobilindustrie weit verbreitete Technologie ist das FMCW-Radar (Frequency Modulated Continuous Wave). Dabei wird ein kontinuierliches Signal ausgesendet, dessen Frequenz während eines sogenannten Chirps linear ansteigt und anschließend wieder abfällt. Diese Modulation erlaubt die gleichzeitige Messung von Entfernung und Geschwindigkeit. Eine schematische Darstellung eines solchen Signals ist in Abbildung 3 zu sehen [3].

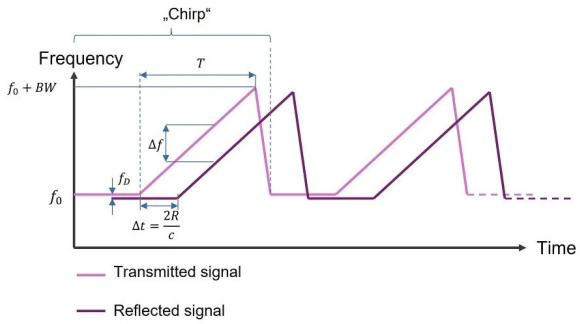


Abb. 2: Frequenzverlauf eines FMCW-Signals [3]

Trifft das Signal auf ein Objekt, wird es mit einer zeitlichen Verzögerung reflektiert. Diese Verzögerung führt zu einer Frequenzdifferenz f zwischen dem ausgesendeten und dem empfangenen Signal. Die Entfernung R zum Objekt ergibt sich daraus wie folgt:

$$R = (c * f * T)/(2 * BW) \quad (1)$$

Dabei steht c für die Lichtgeschwindigkeit im Vakuum, T für die Dauer der Frequenzrampe und BW für die Bandbreite des ausgesendeten Chirp-Signals. Je größer die Frequenzdifferenz, desto weiter ist das reflektierende Objekt entfernt. Befindet sich das Objekt in Bewegung, tritt zusätzlich durch den Dopplereffekt eine Frequenzverschiebung f_D auf. Die daraus resultierende Relativgeschwindigkeit v lässt sich mit folgender Formel bestimmen:

$$v = (c * f_D)/(2 * f_0) \quad (2)$$

Hierbei bezeichnet f_0 die Trägerfrequenz des ausgesendeten FMCW-Signals [3].

Vorgehensweise

Zur Bewertung der Leistungsfähigkeit des Cascade-Radarsensors werden umfangreiche Messreihen unter variierenden Szenarien und Testkonfigurationen durchgeführt. Ziel ist es, die Eignung des Systems zur zuverlässigen Erkennung und Abbildung unterschiedlich geformter Objekte zu untersuchen. Dabei werden sowohl Personen als auch statische Strukturen wie Kartons oder Fensterrahmen erfasst. Der Fokus liegt – wie im Abschnitt zur Analyse des Cascade-Radarsensors beschrieben – auf der detaillierten Auswertung der vom System erzeugten 3D-Punktwolken.

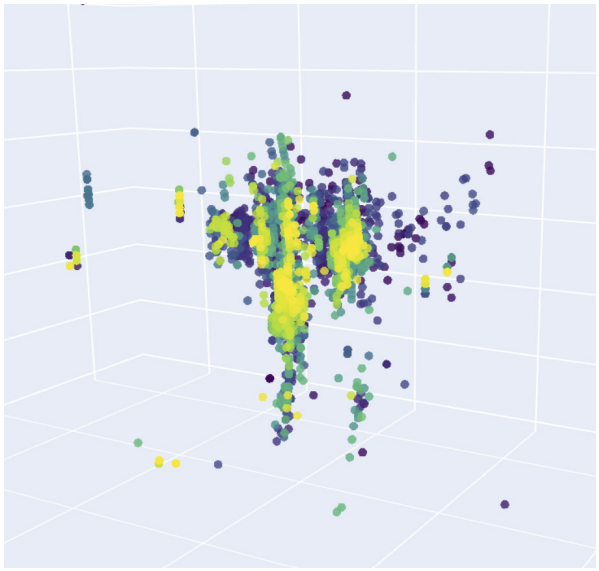


Abb. 3: 3D-Radarpunktwolke einer stehenden Person mit seitlich ausgestreckten Armen [1]

Abbildung 3 zeigt exemplarisch eine Messung, bei der eine Person mit seitlich angehobenen Armen erfasst wurde. Die charakteristische Form ist in der resultierenden Punktwolke deutlich erkennbar. Vergleichbare Messungen werden auch mit weiteren Objekten durchgeführt, um zu evaluieren, wie gut sich das Sensorsystem zur Erkennung unterschiedlicher Konturtypen eignet. Im Mittelpunkt steht dabei nicht nur die grundsätzliche Funktionalität des Radarsensors, sondern insbesondere die Qualität der generierten Punktwolken. Ziel ist es zu prüfen, ob die erfassten Daten hinreichend präzise und strukturiert vorliegen, um perspektivisch als Trainingsbasis für ein neuronales Netz zur konturgetreuen Objekterkennung dienen zu können. Die Bewertung der Radardaten erfolgt anhand definierter Kriterien wie Punktdichte, Formgenauigkeit und Konsistenz.

Einsatzzweck

Radarbasierte Konturenerkennung eröffnet ein breites Spektrum potenzieller Anwendungen. Besonders in Einsatzszenarien, in denen klassische Sensorsysteme durch widrige Umgebungsbedingungen wie Regen, Rauch oder Dunkelheit an ihre Grenzen stoßen, ermöglicht Radar eine robuste und zuverlässige Objekterfassung. Ein exemplarisches Anwendungsfeld stellt die Ausstattung von Feuerwehrkräften dar: In stark verrauchten oder schlecht beleuchteten Umgebungen kann radargestützte Konturenerkennung dazu beitragen, Hindernisse oder Personen gezielt zu lokalisieren und somit die Sicherheit der Einsatzkräfte signifikant zu erhöhen.

Ausblick

Die bisherigen Ergebnisse deuten darauf hin, dass sich der eingesetzte Radarsensor grundsätzlich für die angestrebten Anwendungen zur Konturenerkennung eignet. Eine abschließende Validierung steht jedoch noch aus. In den nächsten Schritten sind umfangreiche Messungen unter verschiedenen Umgebungsbedingungen vorgesehen, um die Leistungsfähigkeit des Systems detailliert zu analysieren und potenzielle Schwachstellen zu identifizieren. Sollten sich die bisherigen Einschätzungen bestätigen, ist eine systematische Aufbereitung der erhobenen Daten für den Einsatz in neuronalen Netzen vorgesehen. In diesem Zusammenhang könnten eigens erstellte Radardatensätze entwickelt und annotiert werden, um Modelle zu trainieren, die spezifisch auf die Anforderungen der konturgetreuen Rekonstruktion abgestimmt sind. Langfristig ergeben sich daraus vielversprechende Einsatzmöglichkeiten in sicherheitskritischen Anwendungsfeldern, in denen konventionelle Sensorik an funktionale Grenzen stößt.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Mordor Intelligence. Radar Sensors Market - Growth, Trends, and Forecasts (2024 - 2029). https://www.mordorintelligence.com/de/industry-reports/radar-sensors-market, 2024.

[3] Marta Martínez Vázquez. The Basics of FMCW Radar. https://www.renesas.com/en/blogs/basics-fmcw-radar, 01 2022.

Entwicklung und Durchführung einer Potentialanalyse für die Konsolidierung von DevOps Toolketten

Nicolai Herrmann **Dennis Grewe**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Andreas Stihl AG & Co. KG, Waiblingen

Einleitung

Durch die ständige Weiterentwicklung von Development Operations (DevOps), was einen Ansatz in der Softwareentwicklung beschreibt, bei dem die Entwicklung (Development) und der IT-Betrieb (Operations) durch Automatisierung und enge Zusammenarbeit zusammengeführt werden, ergeben sich Innovationen. Diese erleichtern die Umsetzung von agilen Entwicklungsmethoden, wodurch der DevOps-Zyklus weiter automatisiert werden soll. Um weiterhin konkurrenzfähig zu bleiben, wird versucht die Softwareproduktentwicklungszeit stetig zu reduzieren und Release Zyklen zu kürzen, etwa anhand von automatisierten Code Qualitätskontrollen durch das Konzept "Continous Integration/Continuous Delivery (CI/CD)". Mithilfe von CI/CD Pipelines können beim Bauen des Codes automatisch verschiedene Test wie Unit Tests, statische Codeanalysen oder End to End Tests ausgeführt und ein erfolgreiches oder nicht erfolgreiches Durchlaufen dieser Tests in einer Übersicht angezeigt werden. So können schnell unterschiedliche Toolketten Verwendung finden und durch steigende Kosten der Anbieter ist eine Evaluation der bereits im Unternehmen STIHL verwendeten DevOps Toolketten in den Fokus gerückt.

Problemstellung

Diese heterogene Tool-Landschaft ist in vielen Unternehmen Realität. Leite et al. [4] zeigen, dass sich in der Praxis häufig inkonsistente Toolchains bilden, die schwer zu integrieren und zu pflegen sind. Dies führt zu organisatorischen Silos, redundanten Anschaffungen und unklaren Verantwortlichkeiten. Besonders problematisch ist dabei, dass ähnliche Funktionen von mehreren Tools abgedeckt werden, ohne dass ein Mehrwert entsteht. So entstehen Kosten für mehrere Programme mit demselben Funktionsumfang. Auch die Zusammenarbeit zwischen den Mitarbeitern wird erschwert, da Expertisen in der Programmbenutzung in den meisten Fällen auf ein Programm beschränkt sind. Falls sich Entwickler in mehreren Programmen auskennen müssen, fallen dementsprechend mehrere Schulungen an, was mehr Kosten und Zeitaufwand bedeutet. Diese Bachelorarbeit führt eine Analyse möglicher Konsolidierungspotentiale der momentan im Unternehmen verwendeten DevOps Programme durch. Ziel ist es durch Analyse der aktuellen Software-Projekte und deren Anforderungen an CI/CD und DevOps Tools eine Vereinheitlichung der Toolkette zu erarbeiten, sodass eine Reduktion der Kosten als auch eine Standardisierung der eingesetzten DevOps Toolketten erreicht und gleichzeitig die Effizienz des Softwareentwicklungsprozesses innerhalb des Unternehmens gesteigert werden kann. Zudem sollen redundante Abdeckungen von Fähigkeiten, die durch mehrere Programme erfolgen, identifiziert werden. Auch gewachsene Abhängigkeiten von Anbietern der Software können festgestellt und gegebenenfalls vermindert oder aufgelöst werden, indem gemäß den Anforderungen der Nutzer in dem Unternehmen entsprechende Alternativprodukte getestet und eingeführt werden. Unter Beachtung von zukunftsorientierten Plänen zur Neuausrichtung in diesem Feld wird eine mögliche Konsolidierung erwogen. Diese Konsolidierungspotentiale sollen unter Anbetracht der aktuell im Unternehmen verwendeten DevOps Toolketten geprüft werden. Eine Untersuchung in multinationalen Unternehmen hat gezeigt, dass heterogene Toolketten in verschiedenen Projekten und Organisationen vergleichbare Herausforderungen bei der Integration und Wartung verursachen, trotz unterschiedlicher eingesetzter Werkzeuge und Konfigurationen. [2] Eine solche Analyse gewährt auch Aufschluss über den Grad der Automatisierung der Softwareentwicklung im Unternehmen und kann ein Vorantreiben von Automatisierungen begünstigen, um den gesamten Softwareentwicklungsprozess effizienter zu gestalten.

DevOps

Für eine belastbare Analyse der einzelnen DevOps Programme und der vorhandenen Toolketten wird sich eng an den einzelnen Schritten des DevOps Zyklus orientiert. Basierend auf Lean- und Agile-Praktiken und einer durchgängigen Automatisierung in der Softwareentwicklung und der Softwarebereitstellung ist das Ziel das Zusammenführen von Entwicklungs- und Betriebsteams sowie deren Aufgabenbereichen. [3] Im Folgenden werden die einzelnen Phasen des DevOps-Zyklus 1 beschrieben und beispielhaft jeweils ein zugehöriges Programm genannt.

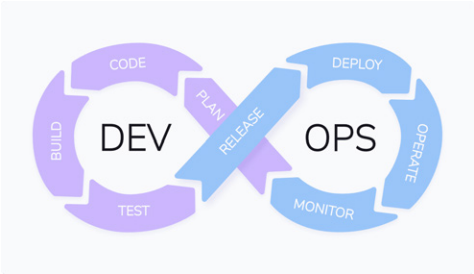


Abb. 1: Abbildung des DevOps Lebenszyklus [5]

In der Planungsphase werden Anforderungen gesammelt, Aufgaben definiert und priorisiert. Ziel ist es, ein gemeinsames Verständnis für das Produkt zu schaffen. Anschließend, in der Entwicklungsphase, schreiben Entwickler Code, führen Unit-Tests durch und committen regelmäßig in ein Versionskontrollsystem. Die Build-Phase beschreibt wie der Quellcode in ausführbare Artefakte umgewandelt wird. Daraufhin kommen in der Testphase automatisierte Tests zum Einsatz, um Funktionalität, Sicherheit und Performance zu überprüfen. Nach bestandener Testphase wird die Anwendung in der Veröffentlichungsphase produktiv freigegeben, zum Beispiel durch automatisierte Release-Prozesse. Die Bereitstellungsphase umfasst die Auslieferung der Anwendung in die Produktionsumgebung. Während des Betriebs wird die Anwendung überwacht, skaliert und gewartet. Abschließend erfolgt das Monitoring, bei dem Nutzerverhalten, System Metriken und Fehler analysiert werden, um Optimierungspotenziale zu identifizieren. [3] In der Praxis werden in den einzelnen Phasen des DevOps-Zyklus zugehörige Programme eingesetzt. Für die Planung eignet sich Jira zur Anforderungsverwaltung. Während der Entwicklung kommen häufig Bitbucket zur Versionskontrolle und GitHub Actions für automatisierte Tests und Integrationen zum Einsatz. Die Build-Phase wird häufig mit einem Packet Manager wie npm und Azure Pipelines umgesetzt. Zur Qualitätssicherung in der Testphase dient unter anderem SonarQube. Für die Bereitstellung kommen automatisierte CI/CD-Pipelines sowie Plattformen wie Kubernetes zum Einsatz. Im laufenden Betrieb unterstützen Grafana zur Visualisierung von Metriken und Azure Monitor zur Systemüberwachung und Analyse.

Ausgangssituation

Zum Startzeitpunkt der Bachelorarbeit waren drei DevOps-Toolketten von verschiedenen Anbietern im Unternehmen in Benutzung. Unterschiedliche Abteilungen verwendeten dabei jeweils eigene Tool Sets für ähnliche Entwicklungsszenarien. Ergänzend kamen weitere Einzelprogramme hinzu, die in die jeweiligen Toolketten integriert wurden. Über einen längeren Zeitraum entstand so ein Flickenteppich an Tools und Prozessen, der sich durch gewachsene Strukturen und parallele Entwicklungen verfestigte. Dies führte zu einer deutlichen Mehrbelastung in Bezug auf Support, Bereitstellung von Software und Infrastruktur, sowie zu einer Kostensteigerung. Der DevOps-Zyklus umfasst die bereits zuvor erwähnten Phasen Planung, Entwicklung, Build, Test, Release, Bereitstellung, Betrieb und Monitoring. Um einen Überblick über die unterschiedlichen Tools der einzelnen DevOps Toolketten von drei großen Anbietern zu bekommen, werden im Anschluss die jeweiligen Tools von Microsoft Azure, GitHub und Atlassian in der folgenden Tabelle 2 gegenübergestellt. Mit Hilfe dieser Gegenüberstellung und weiterer Datenauswertungen können Vereinfachungen und Lücken, sowie Programme mit Konsolidierungspotential erkannt werden. So können Anforderungen der Entwickler unter anderem an ein Versionsverwaltungsprogramm gesammelt und mit den jeweiligen Features der Programme wie GitHub, Bitbucket und Azure Repos verglichen werden. Im Anschluss kann ein Programm ausgewählt werden, das nahezu alle Anforderungen erfüllt und eine Vereinfachung der verwendeten Programme bringen kann.

Feature-Kategorie	Azure DevOps	GitHub	Atlassian
Versionskontrolle	Azure Repos: Git und TFVC	GitHub Repositories: Git	Bitbucket: Git, Mercurial
Projektmanagement	Azure Boards: Scrum, Kanban, Backlogs	GitHub Projects: Kanban, Automatisierungen	Jira Software: Scrum, Kanban, Roadmaps
CI/CD	Azure Pipelines: YAML-Pipelines, Multi-Stage	GitHub Actions: Events, Marketplace	Bamboo: Build- und Deployment-Pläne
Paketverwaltung	Azure Artifacts: npm, NuGet, Maven, Python	GitHub Packages: npm, Maven, NuGet, Docker	Marketplace-Add-ons
Testing	Azure Test Plans: manuell, explorativ, Last-Test	GitHub Actions + Marketplace	Bamboo-Integrationen (z. B. Zephyr)
Sicherheit	Azure Security Center: Policies, Compliance	GitHub Advanced Security: Code Scanning, Secret Detection	Atlassian Access: SAML, 2F
Monitoring	Azure Monitor, Application Insights	GitHub Insights	Opsgenie, Statuspage
Integrationen	Slack, ServiceNow, Azure-Services	Azure, Jira, Slack	REST-API, Marketplace-Apps
Lizenzierung	Pay-as-you-go, User-based	Kostenlos für OSS, zahlend für Teams	Produkt-/Nutzerlizenzen

Abb. 2: Tabellarischer Vergleich unterschiedlicher DevOps Toolketten [1]

Strategie zur Erhebung von Konsolidierungspotentialen

Die Strategie, um Konsolidierungspotential ausfindig zu machen, beinhaltet zwei Schwerpunkte. Ein Schwerpunkt ist die Recherche nach verfügbaren Informationen von Anbietern zu deren DevOps Toolketten sowie technischen Neuerungen und verfügbarer Fachliteratur, um zusätzliche Erkenntnisse aus Studien und Publikationen der Datengrundlage hinzufügen zu können. Um eine Übersicht über das firmeninterne Bild von DevOps zu erlangen, wurde eine interne Umfrage in Bezug auf die verwendeten DevOps Toolketten und Programme gestartet. Dazu wurden in den Bereichen IT und Entwicklung, Abteilungen welche Software entwickeln, ermittelt und die Umfrage an diese weitergeleitet. Dadurch soll abgefragt werden, welche Toolketten und einzelnen Programme bereits verwendet werden, ob es Probleme mit diesen gibt, ob Alternativen zu den bereits genutzten Programmen verwendet werden und welche Anforderungen an Toolketten und Programme gestellt werden. Außerdem wurde die Grundlage für Experteninterviews geschaffen, mit Hilfe derer tiefere Einblicke in die Nutzung und Anforderungen der Anwender erlangt werden sollen. Gemeinsam bilden diese zwei Methoden den zweiten Schwerpunkt. Aus den Interviews und der Umfrage ergibt sich die interne Datengrundlage für die Analyse, was die individuell notwendigen Anforderungen der Anwender sind und ob sich daraus Konsolidierungspotentiale ergeben. Für diese Datengrundlage wird die Umfrage mittels der in Microsoft Forms integrierten Auswertungstools ausgewertet. Die Interviews werden durch die Transkription ebendieser und die Bewertung anhand eines Auswertungskataloges analysiert. Die daraus gewonnenen Erkenntnisse werden aufbereitet und mit der bereits vorhandenen Datengrundlage verglichen, um Übereinstimmungen bei Anforderungen und gebotenen Features sowie eventuellen Fähigkeitslücken zu identifizieren. Daraus können die weiteren Schritte für das Unternehmen abgeleitet werden, wie beispielsweise zwei Programme mit gleichem Funktionsumfang auf eines dieser Programme zu reduzieren um so Kosten zu sparen.

Ausblick

Erste Tendenzen für Konsolidierungen aus den Interviews sind vielversprechend, es deuten sich erste Konsolidierungspotentiale bei Programmen wie der Versionsverwaltung an. Die ausgewerteten Interviews werden mit den bereits recherchierten Daten verglichen und analysiert. Erst nach Auswertung aller Interviews werden die Konsolidierungspotentiale vollständig darstellbar sein. Zusätzlich erfolgt die Auswertung der Umfrage über das Anwendungsverhalten von DevOps Toolketten mit der erste redundante Programme erkannt werden und generelle Toolketten Nutzungen

der einzelnen Abteilungen weiter erschlossen werden sollen. Des Weiteren wird die Datengrundlage aus der Umfrage mit den Ergebnissen der Interviewauswertung verglichen. Das Ergebnis der Auswertung der durch Literaturrecherche, Umfrage und Interviews gewonnenen Daten wird im Anschluss mit den Anforderungen, welche aus der Umfrage hervorgingen, verglichen. Aus diesem Gesamtanalyseergebnis erfolgt die Ableitung einer Handlungsempfehlung bezüglich einer konkreten Konsolidierung der DevOps Toolketten für das Unternehmen STIHL.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Jessica Díaz, Jorge E. Perez, Agustín Yague, Andrea Villegas, and Antonio de Antona. DevOps in Practice – A preliminary Analysis of two Multinational Companies. https://doi.org/10.48550/arXiv.1910.07223, 2019.

[3] Christof Ebert, Gorka Gallardo, Josune Hernantes, and Nicolas Serrano. DevOps. https://doi.org/10.1109/MS.2016.68, 2016.

[4] Leonardo Leite, Carla Rocha, Fabio Kon, Dejan Milojicic, and Paulo Meirelles. A Survey of DevOps Concepts and Challenges. https://doi.org/10.48550/arXiv.1909.05409, 2019.

[5] James Walker. DevOps Lifecycle – The Key Phases of DevOps Workflows. https://spacelift.io/_next/image?url=https%3A%2F%2Fspaceliftio.wpcomstaging.com%2Fwp-content%2Fuploads%2F2025%2F03%2Fwhat-is-devops.png&w=1920&q=75, 2025.

Unveiling the Hidden Threat: Studying Undetected Flaky Test Failures in Large-Scale Continuous Integration Systems

Henrik Herrmann **Dennis Grewe**

Department of Computer Science and Engineering, Esslingen University

Work carried out at BMW AG, München

Introduction

Modern software development relies heavily on Continuous Integration (CI) systems, which aim to ensure that code changes do not introduce unintended negative effects or degrade existing functionality. This article presents a focused exploration of one of the most critical reliability concerns in CI environments: flaky tests. A degradation in functionality caused by a code change is known as a regression, which regression testing aims to identify by regularly executing tests to verify that recent changes do not compromise previously correct behaviour. [2] Flaky tests manifest when they are re-run multiple times, exhibiting a non-deterministic pattern of switching between failing and passing. The intermittent nature of these failures misleads developers, creating false alarms and drawing attention away from genuine issues. The system currently re-runs failed tests immediately in an effort to identify flakiness. If a failed test target passes in a subsequent re-run, it is classified as flaky and typically disregarded as it does not imply a regression. Conversely, test executions that consistently fail are treated as indicators of actual regressions, even though some of them may still be flaky but not exposed during immediate re-running. This misclassification results in considerable time loss and misallocation of effort. Developers often attempt to manually re-run the failed jobs in the pipeline, increasing computational cost. If the error persists even after manual re-runs, they must invest time in searching for the underlying issue, despite the absence of any actual regression. In a large-scale development environment with numerous changes integrated daily, even a small proportion of misclassified test failures accumulates into a substantial productivity burden.

Upgrade Flakiness Detection

The primary goal of this work is to uncover flaky test failures that are misclassified as true regressions and to analyze the effectiveness of current re-running mechanisms. This investigation highlights inefficiencies in current CI strategies and offers enhancements of flakiness detection that we can adopt with minimal disruption to existing workflows. The essence of this work is to introduce delayed validation runs, executed at a later point in time, complementing the immediate re-runs. These delayed runs provide another opportunity to detect the most severe cases of flakiness in test executions that initially fail in all re-runs, offering a more accurate classification of their nature. A key improvement to previous attempts of re-running tests is the systematic approach of re-executing only failed test targets at a periodic interval. Because delayed executions are temporally decoupled from the initial failure, they are more likely to surface intermittent issues, such as a shortage in cloud resources, that are rarely solvable using rapid re-runs. This approach cannot find all flaky tests, but helps increasing the detection rate versus the immediate re-running as well as understanding why flaky failures occur and how effective delayed re-runs are in exposing them. Furthermore, investigations show what the underlying causes are, how we can distinguish different types of flaky tests based on behaviour, and what strategies we should adopt to enhance future CI test reliability. The broader objective is not only to improve test classification but to enable smarter resource allocation regarding developer time and computational cost, ultimately promoting trust in automated testing infrastructures.

Validation Runs

The backbone of the CI pipelines for software development in vehicle dynamics and autonomous driving

systems at BMW consists of two tools. The company uses Bazel for building software, which is a fast, scalable, and extensible open-source build tool to support large codebases with high performance and reproducibility [1]. For CI orchestration, the company employs Zuul, an open-source CI/Continuous Delivery (CD) system designed to handle complex project dependencies and enforce gating workflows to ensure code quality. [6] Examining historical test data from this CI landscape, the focus lies on persistent test failures that occur despite immediate re-running. This dataset of failed test executions includes various projects and subsystems, offering a selected sample of the challenges posed by flaky behaviour in automotive software. A custom script restores the exact repository state at the time of each failure and initiates re-runs using Bazel to rebuild the code and Zuul to manage the repeated test executions. We schedule these executions to run periodically at later intervals, proposedly during nightly validation runs, providing a controlled and consistent environment to reassess previously failed test executions. This process minimizes the influence of external factors and allows for accurate observation in the identical repository condition. As a result, we analyse test outcomes both quantitatively (e.g. Detection Rate Comparison) and qualitatively (e.g. Clustering into Failure Patterns) to determine the frequency and nature of flaky tests, which were previously undetected by immediate re-runs, looking at metrics such as failure reoccurrence and pass-fail transitions, all conducted within the same testing environment.

Expected Improvements

Comparison to other projects at BMW indicates, that many failed tests, which would originally appear as regressions, are actually flaky when re-evaluated through delayed validation runs. This alternative execution strategy significantly improves the detection rate of flaky tests, gaining insight into their characteristics. The advanced knowledge helps developing ways to reduce the time and resources spent on investigating misleading failures. [5] Such strategies include moving flaky tests into the post-submit pipeline, effectively only running them after a change is merged into the main branch. Furthermore, we can adapt flaky test targets to stabilize their behaviour. Root cause analysis identifies various contributing factors, including non-deterministic dependencies such as timing or network issues, limited resources or subtle concurrency bugs that only manifest under specific load conditions or execution orders. [4] Flaky tests differ in their detectability, with some easily exposed by immediate re-runs, others only by delayed validation. If a test consistently fails, it may still be flaky, as establishing a definitive ground truth for a test result is not feasible without access to unlimited testing resources. Understanding these patterns enables the formulation of heuristics that support future predictive models or automated detection systems, similar to those proposed by Hoang and Berding in their Presubmit Rescue approach, which assigns a confidence score indicating the likelihood that a failed test execution is caused by flakiness. [3] These findings underscore the importance of context-aware testing, which adapts to the variability inherent in large-scale software systems. If the strategy for delayed validation runs is not carefully designed, it may lead to increased computational cost or overreliance on imperfect heuristics, therefore introducing new sources of instability rather than resolving existing ones. However, in many cases, the delayed validation runs not only expose flakiness but also offer insights into structural issues in the test code or infrastructure that go unnoticed during routine execution.

Conclusions

The integration of delayed validation runs into BMW's CI pipeline is expected to demonstrate its effectiveness in identifying previously hidden flaky test failures. This enhancement to the testing process will improve the accuracy of test result interpretation, leading to a deeper understanding of behaviour and root causes of flaky test failures. Flaky tests, once seen as an unavoidable nuisance, are better managed through strategic adjustments to test scheduling and result evaluation. As flaky test detection continues to pose challenges, future directions include developing intelligent re-run mechanisms or integrating machine learning models to anticipate flakiness. These tools dynamically adjust re-run strategies based on test history and system load, further refining the balance between speed and reliability. Ultimately, this work emphasizes the need to regard test reliability as a dynamic attribute that benefits from continuous observation, analysis, and refinement. By enhancing the testing pipeline with targeted validation executions, organizations can significantly improve CI system efficiency, reducing the consumption of computing resources, minimizing unnecessary developer debugging effort, and promoting more reliable test result interpretation. These improvements support smarter allocation of time and infrastructure, ensuring that engineering efforts address true regressions rather than chasing phantom failures.

References and figures

[1] Project Contributors Bazel. Intro to Bazel. https://bazel.build/about/intro, 2025.

[2] Renan Greca, Breno Miranda, and Antonia Bertolino. Orchestration strategies for regression test suites. In *2023 IEEE/ACM International Conference on Automation of Software Test (AST)*, pages 163–167. IEEE, 2023.

[3] Minh Hoang and Adrian Berding. Presubmit Rescue: Automatically Ignoring Flaky Test Executions. In *Proceedings of the 1st International Workshop on Flaky Tests*. Association for Computing Machinery, 2024.

[4] Qingzhou Luo, Farah Hariri, Lamyaa Eloussi, and Darko Marinov. An empirical analysis of flaky tests. In *Proceedings of the 22nd ACM SIGSOFT international symposium on foundations of software engineering*, pages 643–653. Association for Computing Machinery, 2014.

[5] Owain Parry, Gregory M Kapfhammer, Michael Hilton, and Phil McMinn. A survey of flaky tests. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 31:1–74, 2021.

[6] Project Contributors Zuul. About Zuul. https://zuul-ci.org/docs/zuul/latest/about.html, 2012.

Erstellung und Entwicklung eines Monitoring-Konzepts für eine Azure-basierte Integrationslösung

Arne Hobrlant **Jörg Nitzsche**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Mercedes-Benz Mobility AG, Stuttgart

Einleitung

Durch die fortschreitende Digitalisierung werden immer mehr IT-Systeme miteinander verbunden, um unter anderem effizientere Geschäftsprozesse zu ermöglichen [3]. Integrationslösungen spielen eine zentrale Rolle, da die Kommunikation und Interaktion zwischen unterschiedlichen Systemen aufgrund technologischer Unterschiede oft nicht selbstverständlich sind. Die primäre Aufgabe einer Integrationslösung ist daher, einen reibungslosen Prozess zwischen zwei oder mehreren Schnittstellen zu gewährleisten. Dabei ist das Überwachen der laufenden Prozesse innerhalb der Integrationslösung entscheidend, um eine stabile Interaktion zwischen den Systemen sicherzustellen und potenzielle Fehler frühzeitig zu erkennen und zu beheben. Damit dies möglich ist, müssen die Verantwortlichen der Integrationsplattform einen Überblick über alle relevanten Prozesse haben sowie über ein zuverlässiges Benachrichtigungssystem verfügen, das bei bestimmten Ereignissen sofort informiert.

Ziel des Projekts

Ziel dieser Arbeit ist es, für eine selbst entwickelte Integrationsplattform ein Konzept zu erstellen und umzusetzen, das sämtliche Prozesse überwacht. Die Integrationsplattform wird entwickelt, um die Kommunikation zwischen verschiedenen Märkten im Rahmen einer weiteren Mercedes-Benz Mobility Softwarelösung zu ermöglichen (siehe Abbildung 1). Sie erfüllt dabei zwei zentrale Aufgaben:

1. Schnittstellenkommunikation: Die Plattform stellt Verbindungen zwischen unterschiedlichen Systemen her, die auf verschiedenen Technologien basieren. Daher ist sie in der Lage, Anfragen und Antworten in unterschiedlichen Formaten entgegenzunehmen, zu verarbeiten und weiterzuleiten.
2. Datenanpassung: Anfragen und Antworten müssen vor der Weitergabe inhaltlich angepasst werden, da die empfangenen Daten nicht im Originalformat an die Zielsysteme übermittelt werden können. Die Plattform übernimmt somit die Transformation der Daten, um deren Verarbeitung durch das jeweils angeschlossene System sicherzustellen.

Um die Stabilität und Zuverlässigkeit der Plattform zu gewährleisten, muss ein umfassendes und sicheres Überwachungskonzept eingebunden werden. Durch diese Überwachung soll zum einen die erfolgreiche Ausführung der Prozesse bestätigt werden. Zum anderen soll sichergestellt werden, dass im Fehlerfall die Verantwortlichen umgehend informiert werden, damit die Probleme schnellstmöglich behoben werden können. Das Monitoring soll dabei sowohl die einzelnen Prozessabläufe als auch technische Defekte der Plattform selbst erfassen und überwachen.

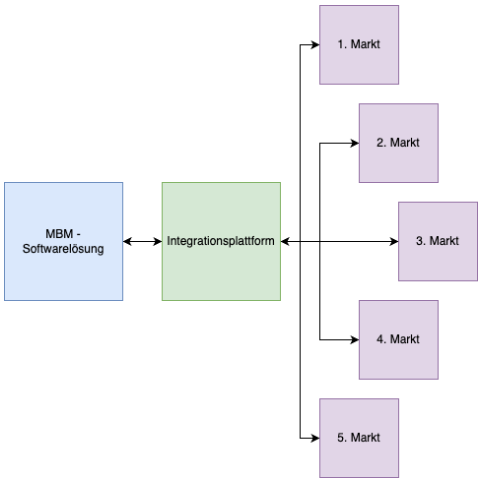


Abb. 1: Struktur der Integrationslösung zur Marktkommunikation [1]

Datadog

Im Verlauf dieser Arbeit wird Datadog als zentrales Werkzeug zur Prozessüberwachung eingesetzt. Datadog ist eine cloudbasierte Monitoring- und Analyseplattform, die es ermöglicht, sowohl Anwendungs- als auch Infrastrukturmetriken in Echtzeit zu erfassen, zu visualisieren und Alarme zu konfigurieren [2]. Dadurch eignet sich Datadog ideal, um komplexe Prozessabläufe und Systemzustände zentral zu überwachen und bei auftretenden Fehlern schnell zu reagieren. Die Integrationsplattform speichert alle Log-Einträge in Datadog, sodass eine detaillierte Nachverfolgung der einzelnen Unterprozesse möglich ist und eine effiziente Fehleranalyse sowie proaktive Benachrichtigung der Verantwortlichen gewährleistet werden kann.

Vorgehensweise

Die Vorgehensweise des Projektes erfolgt über zwei primäre Schritte.

1. Anforderungsanalyse
2. Technische Umsetzung

Die Anforderungsanalyse ist das Fundament der Arbeit, da hier definiert wird, welche Prozesse wie überwacht werden sollen. Zu diesem Zweck werden Meetings mit dem verantwortlichen Fachbereich und technischen Ansprechpartnern durchgeführt. In diesen Besprechungen werden die überwachenden Systemprozesse identifiziert, in Unterprozesse aufgeteilt und priorisiert. Zudem wird festgelegt, welche Kriterien für eine erfolgreiche oder fehlerhafte Ausführung gelten und welche Benachrichtigungsmechanismen im Falle eines Fehlers greifen sollen. Alle Anforderungen werden dokumentiert und dienen im weiteren Verlauf als Basis für die technische Umsetzung im Monitoring-System.
Auf Basis der definierten Anforderungen wird die technische Umsetzung in Datadog durchgeführt. Zunächst werden die identifizierten Prozesse so angepasst, dass sie aussagekräftige Log-Einträge erzeugen, die anschließend in Datadog gespeichert und verarbeitet werden. Daraufhin wird ein Dashboard erstellt, das den Status der Unterprozesse visualisiert. Erfolgreiche Abläufe werden dabei grün, fehlgeschlagene rot dargestellt (siehe Abbildung 2). Zusätzlich werden Monitore konfiguriert, die sowohl ereignisgesteuerte als auch zeitbasierte Alarme auslösen. So kann z. B. eine Benachrichtigung erfolgen, wenn ein bestimmter Prozess fehlschlägt oder innerhalb der letzten 24 Stunden nicht ausgeführt wurde. Durch diese zweistufige Vorgehensweise wird sichergestellt, dass die Überwachung der Prozesse sowohl inhaltlich korrekt als auch technisch effizient umgesetzt wird.

Abb. 2: Beispiel einer Übersicht im Datadog Dashboard [1]

Ausblick

Durch die Arbeit entsteht eine Überwachungslösung, welche einen wichtigen Aspekt der Sicherstellung von Zuverlässigkeit innerhalb der Integrationsplattform abdeckt. Ein stabiles Überwachungs- und Benachrichtigungssystem kann mit der Integration von Datadog erreicht werden, das sowohl technische als auch prozessuale Fehler frühzeitig erkennt und meldet. Es gibt einige Erweiterungsmöglichkeiten für die Zukunft. Man kann das gesamte Monitoring durch die Aufnahme weiterer Prozesse erweitern, um eine noch umfassendere Überwachung des Systems zu ermöglichen. Des Weiteren ist es möglich, automatisierte Reaktionen auf bestimmte Events zu implementieren, wie etwa einen selbstheilenden Mechanismus zu starten oder die Anbindung eines Incident-Management-Systems wie ServiceNow. Langfristig könnte auch eine Auswertung der gesammelten Monitoring-Daten im Sinne eines proaktiven Prozess- und Qualitätsmanagements erfolgen. Dadurch lassen sich dann möglicherweise wiederkehrende Fehlerquellen identifizieren, Trends analysieren und Optimierungspotenziale erkennen. Insgesamt legt dieses Projekt die technische und konzeptionelle Grundlage für ein skalierbares und zukunftsfähiges Überwachungssystem, das einen wertvollen Beitrag zur Betriebssicherheit und Effizienz der Integrationsplattform leistet.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Inc. Datadog. Monitors. https://docs.datadoghq.com/monitors/, 2025.

[3] Andreas Streim. Unternehmen treiben mit der Cloud ihre Digitalisierung voran. https://www.bitkom.org/Presse/Presseinformation/Unternehmen-treiben-mit-Cloud-Digitalisierung-voran?utm_source=chatgpt.com, 2024.

Next-Gen Crash Analysis: Machine Learning Surrogates for FEM Simulations

Dieter Holstein **Steffen Schober**

Department of Computer Science and Engineering, Esslingen University

Work carried out at Dr. Ing. h.c. F. Porsche AG, Weissach

Motivation and Problem Definition

Vehicles are essential in today's world, offering efficient transportation that enables faster and farther travel. However, the rising number of vehicles has led to increased risks of accidents and injuries. The World Health Organization [6] reported that in 2023 alone, approximately 1.19 million people died in road traffic accidents. [5] In response, governments worldwide have enacted safety regulations to mitigate accident risks and enhance road safety. [4] These regulations ensure that vehicles include various safety features to protect occupants. One key requirement is the US standard FMVSS 201U, which aims to minimize head injury risks, measured by the Head Injury Criterion (HIC). [8] This standard mandates that vehicles sold in the US must ensure that collisions in the upper vehicle interior do not exceed a HIC of 1000, as determined by tests involving a Free Motion Headform (FMH) launched at specific interior locations at a speed of 15 mph. [2]

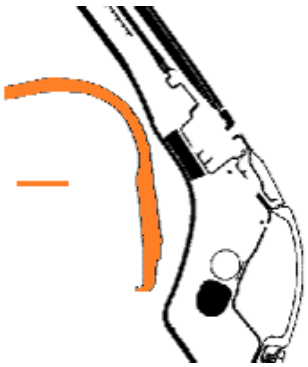


Fig. 1: Vertical cross-section of the vehicle's roof with the FMH in orange [7]

During the initial development phase, experts design structural components while considering cost, weight, design targets, and the implementation of safety features to meet regulations. They create two-dimensional sketches to visualize structural layouts and discuss variations. To better illustrate how these sketches might look like, Figure 1 depicts a section of the roof from a vertical cross-section of the vehicle.
This cross-section, which is perpendicular to the shot location, illustrates the vehicle's structural components by depicting their surface area when bisected, with the FMH shown in orange on the left.
However, there are only very limited possibilities to evaluate these drafts at this early stage, and most decisions are based purely on experience, without any guarantee that the design will truly meet the defined HIC value. If the design progresses further and later simulations reveal that the HIC criterion is not met, all stages up to the initial design stage would need to be redone, resulting in high costs and long development cycles.
To address this challenge, methods are needed that provide earlier feedback and better support the decision-making during the concept phase. Instead of relying solely on simulations at a later stage, integrating data-driven insights into the traditionally experience-based process could enable a more informed and proactive design workflow.
In this context, the goal of this work is to explore the potential of machine learning techniques to enhance early-stage development by generating sequential frames from initial design sketches. These frames are intended to visually illustrate the dynamic interaction between the FMH and interior components over time, effectively forming an animation. Such a visual representation could help identify critical issues earlier in the process, reduce the number of testing and development cycles, and enable more targeted optimization of design ideas while ensuring compliance with safety standards.

Methodology

To address the overarching goal of this Master's thesis, a structured and systematic approach will be adopted.

The following methodology outlines a clear roadmap that guides the research and evaluation phases.

Research:
The thesis will begin with a comprehensive review of existing methods for image and video generation, with a particular focus on techniques that incorporate physical properties during the generation process. Several architectures and approaches will be examined, analyzing their strengths, limitations, and suitability for the given problem.

Architectural Overview:
Based on the research findings, the next step will involve establishing an architectural overview of current image generation techniques that could be applied to the targeted problem. Particular attention will be given to approaches capable of generating sequential frames that take into account the physical properties and material deformation during the impact of the FMH. The identified methods will be compared and evaluated in terms of their applicability and performance potential.

Data Preparation:
Following the architectural overview, the focus will shift to preparing suitable training data. The basic cross-section data is already available in multiple variations. In this phase, an appropriate data representation will be selected, and suitable data splits and training strategies will be defined based on the selected models and insights gained from the research.

Implementation:
Subsequently, selected approaches will be implemented and evaluated based on the insights gathered in the previous phases. Multiple evaluation criteria will be defined to benchmark the models' performance. Rather than delivering a fully refined and operational tool, the primary objective is to create a proof of concept. This proof of concept will assess the feasibility of applying machine learning methods in the early stages of vehicle development based on initial two-dimensional sketches, while also considering the requirements of the US safety standard FMVSS 201U. [8]

Model trustworthiness:
As the automotive crash domain is inherently safety-critical, the interpretability and transparency of the proposed machine learning model could be taken into account. Therefore, this thesis might briefly explore the concepts of:

- **Explainable AI:** Methods that aim to make model predictions understandable to humans. [3]
- **Physics-Informed Neural Networks:** Neural networks that incorporate physical laws directly into the training process. [1]

While model trustworthiness may play a relevant role in safety-related applications, its treatment in this work will be relatively limited, primarily serving as a supporting lens for assessing the reliability and decision-making processes of the selected methods.

References and figures

[1] Salvatore Cuomo et al. Scientific Machine Learning Through Physics–Informed Neural Networks: Where we are and What's Next. https://doi.org/10.1007/s10915-022-01939-z, 2022.

[2] National Highway Traffic Safety Administration Department of Transportation. Federal Motor Vehicle Safety Standard. 49 CFR 571.201 - Standard No. 201; Occupant protection in interior impact. https://www.ecfr.gov/current/title-49/subtitle-B/chapter-V/part-571/subpart-B/section-571.201, 2004.

[3] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable AI: A Review of Machine Learning Interpretability Methods. https://www.mdpi.com/1099-4300/23/1/18, 2020.

[4] World Health Organization. Strengthening road safety legislation: a practice and resource manual for countries. https://iris.who.int/handle/10665/85396, 2013.

[5] World Health Organization. Global status report on road safety 2023. https://www.who.int/publications/i/item/9789240086517, 2023.

[6] World Health Organization. World Health Organization. https://www.who.int, 2025.

[7] Own representation.

[8] Helen A. Rychlewski, Jessica A. Cronkhite, and Michael J. Smith. FMVSS 201U Testing - Vehicle Targeting Using both Manual and Computer-Aided Methods. https://www.sae.org/content/1999-01-0434/, 1999.

Der globale Wettbewerb um Halbleiter: Technologische Analyse der Halbleiterindustrie, wirtschaftliche Auswirkungen von US-Exportkontrollen und Chinas Strategien zur Entwicklung eigener Fertigungskapazitäten

Nasrullah Idkhafif **Giles-Arnaud Nzouankeu Nana**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

Halbleiter stellen eine Schlüsseltechnologie dar, die in nahezu allen modernen elektronischen Systemen zum Einsatz kommt – von Consumer Electronics über industrielle Steuerungssysteme bis hin zu sicherheitskritischen Anwendungen. Ihre technologische Relevanz ist eng mit ihrer komplexen Herstellung verbunden, bei der präzise physikalische und chemische Prozesse aufeinander abgestimmt werden müssen.
Die Fertigung hochintegrierter Schaltkreise erfordert spezialisierte Produktionsanlagen, hochreine Materialien sowie umfassendes Know-how in mehreren Disziplinen – von der Nanotechnologie bis zur Software-basierten Prozesssteuerung. Diese hohe technologische Komplexität begrenzt die Zahl der Akteure, die in der Lage sind, leistungsfähige Halbleiter selbst zu entwickeln und zu produzieren. [1]
Halbleiter stellen nicht nur eine technische Ressource dar, sondern auch ein Beispiel für systemisch vernetzte Lieferketten, bei denen digitale Planung, Simulation und Automatisierung entscheidende Rollen spielen. Gleichzeitig zeigen sich durch die globale Konzentration von Fertigungskapazitäten Abhängigkeiten, die sowohl ökonomische als auch strategische Implikationen besitzen. [2]
In diesem Kontext gewinnen Maßnahmen wie Exportkontrollen oder nationale Förderprogramme an Bedeutung. Sie reflektieren nicht nur geopolitische Spannungen, sondern auch das Bewusstsein für die strategische Relevanz einer stabilen und zugänglichen Chipversorgung.
Die vorliegende Arbeit untersucht die technologischen Grundlagen der Halbleiterproduktion und analysiert, welche wirtschaftlichen Effekte durch US-Exportbeschränkungen entstehen. Zudem wird geprüft, wie China versucht, eigene Fertigungsressourcen aufzubauen und welche Herausforderungen dabei auftreten. Ziel ist es, die Wechselwirkungen zwischen Technologieentwicklung, globaler Lieferkettenlogistik und staatlichem Handeln aus einer informationswirtschaftlichen Sichtweise heraus zu beleuchten. [4]

Halbleiterfertigungsprozess

Halbleiter sind Schlüsselkomponenten moderner Elektronik und spielen eine zentrale Rolle in Anwendungen, die vom Smartphone über Automotive bis hin zu medizinischen Geräten reichen. Die Herstellung hochintegrierter Chips ist ein äußerst komplexer Prozess, der aus einer Vielzahl von Einzelschritten besteht, die jeweils hohe Präzision, Reinheit und Kontrolle erfordern.
Die Produktion beginnt mit der Züchtung eines monokristallinen Silizium-Ingots, meist durch das Czochralski-Verfahren. Dieser wird anschließend in dünne Scheiben – sogenannte wafers – geschnitten und poliert, um eine glatte Oberfläche für die nachfolgenden Strukturierungsprozesse bereitzustellen. Diese Wafer bilden die Grundlage für die darauf ablaufenden lithografischen und chemischen Verfahren.

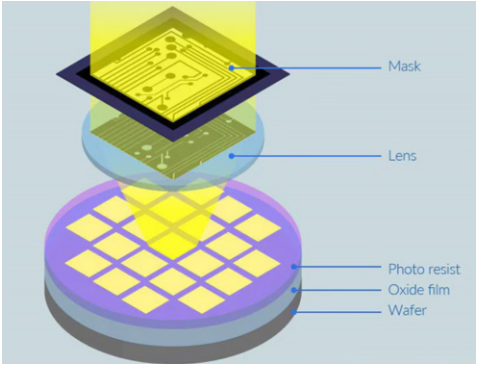


Abb. 1: Veranschaulichung der Photolithografie [3]

Ein zentraler Schritt im Fertigungsprozess ist die photolithography . Sie ermöglicht die Übertragung mikroskopisch kleiner Schaltkreismuster auf die Waferoberfläche und ist damit essentiell für die Erzeugung der elektrischen Bauelemente wie Transistoren und Leiterbahnen.
Vor der Belichtung wird der Wafer zunächst mit einem lichtempfindlichen Material, dem photoresist , beschichtet. Dies erfolgt üblicherweise durch spin coating , bei dem der Wafer rotiert und das Photoresist gleichmäßig verteilt wird. Es gibt zwei Arten von Photoresists: positive resists , bei denen die belichteten Bereiche im Entwicklungsprozess gelöst werden, und negative resists , bei denen die ungebelichteten Bereiche entfernt werden.

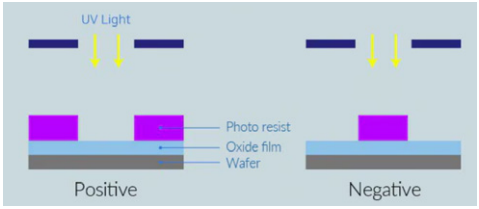


Abb. 2: Abbildung des Photoresist [3]

Nach dem Auftrag des Photoresists folgt die mask alignment -Phase. Dabei wird eine photomask – eine Glas- oder Quarzplatte mit einem Muster aus lichtundurchlässigem Material wie Chrom – millimetergenau über dem Wafer positioniert. Mit Hilfe eines stepper - oder scanner -Systems wird sichergestellt, dass das Muster korrekt übertragen wird.
Im nächsten Schritt erfolgt die exposure des Wafers mit ultraviolettem Licht. In modernen Fertigungslinien kommt hierbei zunehmend EUV-Technologie (extreme ultraviolet) zum Einsatz, da sie deutlich feinere Strukturen ermöglicht als herkömmliche Verfahren.
Nach der Belichtung wird der Wafer in eine developer -Lösung getaucht, welche die löslichen Bereiche des Photoresists entfernt. Bei positivem Resist bleiben dadurch die nicht belichteten Stellen erhalten, während beim negativen Resist die belichteten Bereiche stabil bleiben. Nach dem development -Prozess wird der Wafer gründlich gespült und getrocknet, sodass nur noch die gewünschten Muster sichtbar sind.
Diese strukturierten Bereiche dienen anschließend als Maskierung für nachfolgende Prozesse wie etching oder deposition . Beim dry etching wird mithilfe von Plasma selektiv Material entfernt, um dreidimensionale Strukturen wie Transistoren oder interconnects herzustellen. Alternativ kommt wet etching zum Einsatz, bei dem chemische Lösungen verwendet werden.
Weitere kritische Schritte umfassen die gezielte Dotierung des Halbleitermaterials durch ion implantation oder diffusion , um n- und p-dotierte Bereiche zu erzeugen. Danach erfolgt die metallization , bei der leitfähige Schichten wie Kupfer oder Aluminium abgeschieden werden, um die elektrischen Verbindungen zwischen den einzelnen Komponenten herzustellen.
Zum Abschluss erhält der Chip eine passivation layer , oft aus Siliziumnitrid oder Siliziumdioxid, die vor Umwelteinflüssen schützt. Im letzten Schritt, dem packaging , wird der Wafer in Einzelchips getrennt (dicing), montiert und in ein Gehäuse eingebettet, das sowohl mechanischen Schutz als auch die elektrische Anbindung an externe Systeme gewährleistet.
Die kontinuierliche Weiterentwicklung dieser Prozesse, insbesondere in der photolithography , ist entscheidend für die Fortschritte in der Halbleiterindustrie. Die Einführung von EUV-Lithografie und innovativen Packaging-Technologien zeigt, wie eng Forschung, Automatisierung und digitale Steuerung in der Chipproduktion verzahnt sind.
Gleichzeitig verdeutlicht die hohe technologische Komplexität der Fertigung, warum diese Prozesse nur von einer begrenzten Zahl von Unternehmen und Ländern beherrscht werden. Die Abhängigkeit von speziellen Maschinen, Rohmaterialien und Know-how führt zu systemischen Risiken entlang der Lieferketten und macht Halbleiter zu einem zentralen Thema in der internationalen Technologiepolitik. [3] [1] [2]

US-Exportkontrollen

Seit Oktober 2022 hat die US-Regierung umfassende Exportbeschränkungen für Halbleiter, Computersysteme mit diesen Komponenten sowie für Fertigungsanlagen eingeführt, die nach China exportiert werden dürfen. Die Maßnahmen zielen darauf ab, Chinas Zugang zu fortschrittlichen Technologien einzuschränken, insbesondere in Bereichen wie Künstlicher Intelligenz und Supercomputing. Ziel ist es, sowohl den Import hochentwickelter Chips als auch den Aufbau eigener Produktionskapazitäten durch chinesische Unternehmen zu erschweren.
Die Exportkontrollen wurden im Laufe der Zeit weiter verschärft. Im Jahr 2023 und 2024 folgten Nachbesserungen, die den Druck auf die chinesische Halbleiterindustrie verstärkten. Zudem haben einige Verbündete der USA ähnliche Beschränkungen übernommen, was die internationale Wirkung dieser Maßnahmen verstärkte. Auch nach einem Regierungswechsel setzte die neue Administration diese Politik fort und erweiterte sie sogar durch zusätzliche Exportverbote gegenüber chinesischen Unternehmen.
Die unmittelbaren Folgen waren deutlich spürbar: Chinesische Akteure mussten sich schnell anpassen, da Lieferengpässe entstanden und Preise für bestimmte Bauelemente stiegen. Einige Unternehmen reduzierten ihre Belegschaften oder passten ihre Geschäftsmodelle an. Gleichzeitig führten die Kontrollen jedoch auch dazu, dass chinesische Unternehmen verstärkt in

die Eigenentwicklung investierten, unterstützt von staatlichen Förderprogrammen.
Ein prominentes Beispiel dafür ist Huawei. Nachdem das Unternehmen 2019 vom Zugang zu US-Technologie abgeschnitten wurde, begann es eine umfassende Strategie zur Unabhängigkeit von ausländischen Lieferketten. Bis 2024 gelang es Huawei, wieder Produkte mit leistungsfähigen Halbleitern auf den Markt zu bringen – darunter auch Entwicklungen im Bereich der künstlichen Intelligenz. Dies zeigt, dass Exportbeschränkungen zwar kurzfristige Wirkung zeigen können, langfristig aber Innovation und Selbstständigkeit antreiben können.
Trotz der strengen Kontrollen bleiben jedoch Lücken: Während es bei schwer transportierbarem Equipment wie Fertigungsanlagen relativ einfach ist, Exporte zu regulieren, sind Halbleiter selbst leicht versteckbar und in großer Stückzahl produziert. Ebenso lässt sich Software, beispielsweise Design-Tools, kaum vollständig kontrollieren. Es gibt bereits Berichte über den Einsatz von Scheinfirmen, um Exportverbote zu umgehen – ein Indiz dafür, dass solche Maßnahmen nicht uneingeschränkt durchsetzbar sind.
China reagiert auf diese Herausforderungen mit einer systematischen Stärkung der eigenen Industrie. Der Aufbau nationaler Kapazitäten in Design, Fertigung und Verpackung wird aktiv gefördert, mit dem Ziel, langfristig unabhängig von westlichen Lieferketten zu sein. Diese Entwicklung unterstreicht, dass der globale Wettbewerb um Halbleiter nicht nur technologisch, sondern auch strategisch entschieden wird.
Die hier dargestellten Entwicklungen verdeutlichen nur einen Ausschnitt des komplexen Zusammenspiels zwischen Technologie, Handelspolitik und strategischem Wettbewerb. Der globale Halbleitermarkt ist von tiefgreifenden Abhängigkeiten, technischen Herausforderungen und politischen Entscheidungen geprägt. Es gibt zahlreiche weitere Aspekte – von der Rohstoffversorgung über Lieferkettenresilienz bis hin zur internationalen Kooperation –, die bei der Bewertung zukünftiger Entwicklungen berücksichtigt werden sollten. [4]

Literatur und Abbildungen

[1] Venus Kohli. Semiconductor Fabrication Process: The Ultimate Guide to Creating Cutting-Edge Electronics. https://www.wevolver.com/article/semiconductor-fabrication-process-the-ultimate-guide-to-creating-cutting-edge-electronics, 04 2023.

[2] Gary May and Costas Spanos. *Fundamentals of Semiconductor Manufacturing and Process Control*. John Wiley & Sons, Inc., 2006.

[3] Semiconductor Europa GmbH Samsung et al. Part 4, Drawing Structures in Nano-scale. https://semiconductor.samsung.com/emea/support/tools-resources/fabrication-process/eight-essential-semiconductor-fabrication-processes-part-4-photolithography-laying-the-blueprint/, 09 2017.

[4] Charles Wessner, Sujai Shivakumar, and Thomas Howell. The Limits of Chip Export Controls in Meeting the Challenge of China. https://www.csis.org/analysis/limits-chip-export-controls-meeting-china-challenge, 04 2025.

Enhancing Text-Based Object Detection Models for Industrial Applications

Manuel Kaiser **Dieter Morgenroth**

Department of Computer Science and Engineering, Esslingen University

Work carried out at Sick AG, Hamburg

Introduction

In modern industrial settings, automatic image processing plays a crucial role in ensuring efficient and accurate production and logistics. Text-based object detection models offer a promising alternative to traditional category-based methods. Instead of being trained on fixed class labels, these models use natural language prompts to detect objects, which makes them open-set capable. This means they are designed to detect objects that were not explicitly included as specific classes during training. However, the current generation of text-image models is not well adapted to the specialized language and visual characteristics found in industrial environments. This is largely because they are trained on large-scale, general-purpose datasets such as COCO [2]. This mismatch introduces several challenges. First, the language in these datasets often lacks domain-specific terms, which leads models to misinterpret technical object names or component descriptions. Second, industrial images usually differ from everyday scenes in structure, lighting, and object presentation. As a result, models trained on general data may miss small or occluded components or confuse similar-looking object parts. Finally, high-quality industrial datasets with both segmentation masks and matching textual descriptions are rare, creating a bottleneck for supervised training or fine-tuning. To address these limitations, this thesis explores whether adapting state-of-the-art text-based detection models can improve their performance in industrial contexts. The focus lies on fine-tuning a model using a domain-specific dataset and investigating scalable ways to generate image-text data using automated captioning tools.

Background and Related Work

The foundation of modern vision-language models lies in the Transformer architecture, introduced by Vaswani et al. [6], which revolutionized natural language processing and has since been extended to vision tasks [1]. This architecture enables multimodal learning by projecting text and image inputs into a shared embedding space, allowing flexible interactions between the modalities. A number of recent models have taken advantage of this capability to perform open-vocabulary object detection. Among them, Grounding DINO [3] stands out as a model that integrates a text encoder with a Transformer-based image encoder, enabling it to match flexible text queries with corresponding regions in the image. The model leverages cross-attention mechanisms to align visual features with text embeddings, supporting zero-shot detection of objects described in natural language. The architecture is shown in detail in the Figure 1. Other models such as OWLv2 [4] and YOLOE [7] also tackle open-set detection tasks and will be examined alongside Grounding DINO in this thesis.

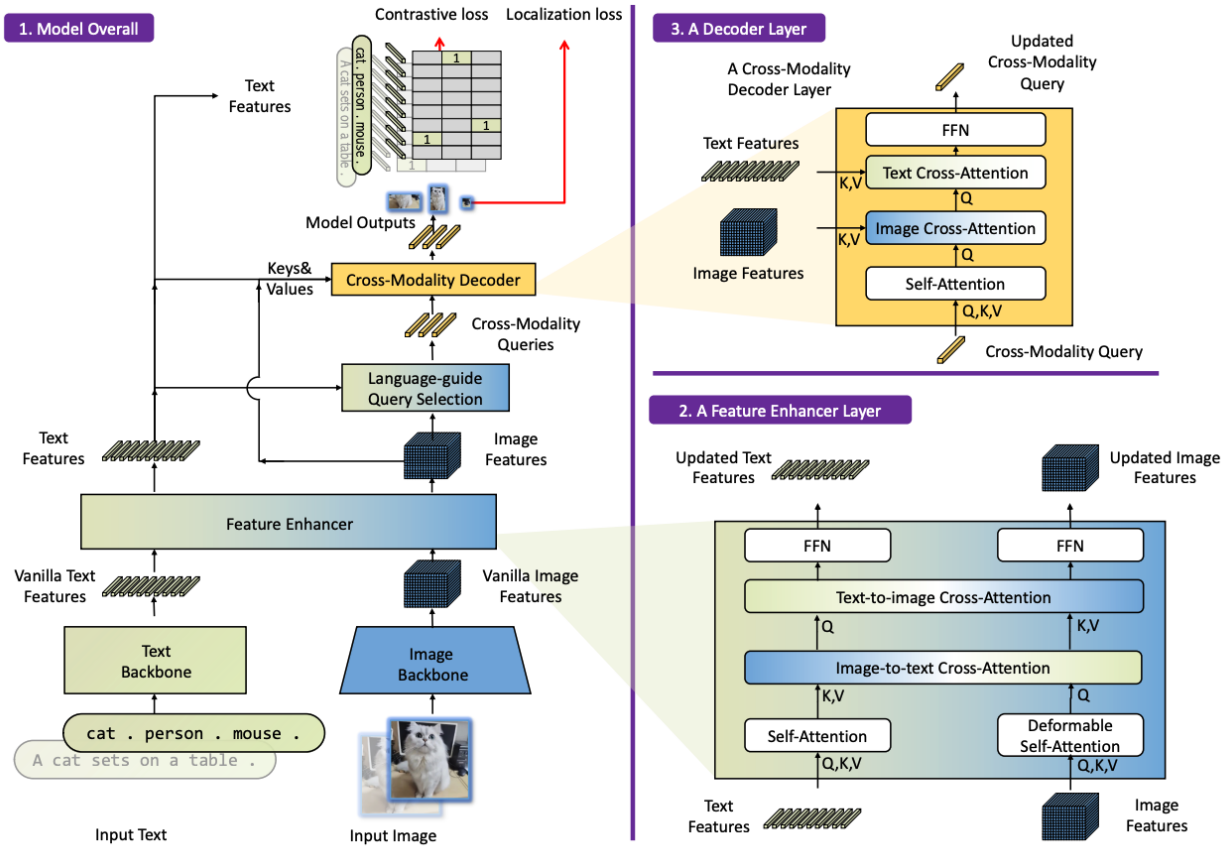


Fig. 1: GroundingDINO Structure [3]

Methodology

First, a model analysis is conducted by evaluating the baseline performance of the models on a small industrial dataset. This step aims to identify both the strengths and the typical failure modes in detection and grounding behavior. Next, two strategies are pursued for dataset generation. In the manual approach, a small but high-quality dataset is created by manually labeling a limited number of industrial images with accurate and consistent captions. In the automated approach, a large-scale dataset is generated using image captioning models to automatically produce descriptive labels for a broad set of industrial images. Following this, the entire model is fine-tuned using the curated dataset to improve its performance on domain-specific tasks. Finally, in the evaluation phase, the performance of the fine-tuned model is benchmarked against its original version using standard detection metrics.

Baseline Results

In the first experiment a ground truth dataset with 750 images has been used to evaluate the performance of these base models. For every image, the following textual input was used: "cardboard. bag. cylinder. bin container." Afterwards, all the predictions from the models have been extracted and compared to the ground truth.

For a prediction to be counted as a true positive the label must be correct and the intersection of union is supposed to be 0.5 or above. Table 2 shows the results of this experiment. The results indicate that all of the models perform poorly overall and show considerable variation in the specific classes they struggle with. An example of a ground truth image paired with the respective model prediction is displayed in Figure 3.

Class	Grounding-DINO	OwlV2	YOLOE
cardboard	0.535	0.495	0.044
bag	0.265	0.010	0.316
cylinder	0.013	0.345	0.365
bin container	0.53	0.738	0.656

Fig. 2: F1-Score for different models on same Text Prompt using Evaluation Dataset [5]

Outlook

The performance of vision-language models is highly dependent on the nature and quality of the data they are trained on. Consequently, one key focus of this thesis is to investigate how fine-tuning with domain-specific datasets can improve detection performance in industrial environments. To this end, various dataset compositions will be explored, ranging from small, manually labeled datasets with carefully curated captions to large-scale datasets generated using automated methods. This exploration aims to determine what types of annotations are most beneficial for improving the model's ability to detect and localize industrial components based on textual prompts. In particular, the trade-off between annotation effort and performance gain will be assessed to identify scalable strategies for real-world deployment. Furthermore, the feasibility of employing image captioning tools to automatically generate descriptive labels will be examined. These tools could offer a practical solution for overcoming the scarcity of annotated industrial data by enabling the creation of large training datasets with minimal manual intervention. By evaluating both manual and automated approaches, this thesis aims to provide insights into effective data strategies for adapting multimodal object detection models to specialized domains.

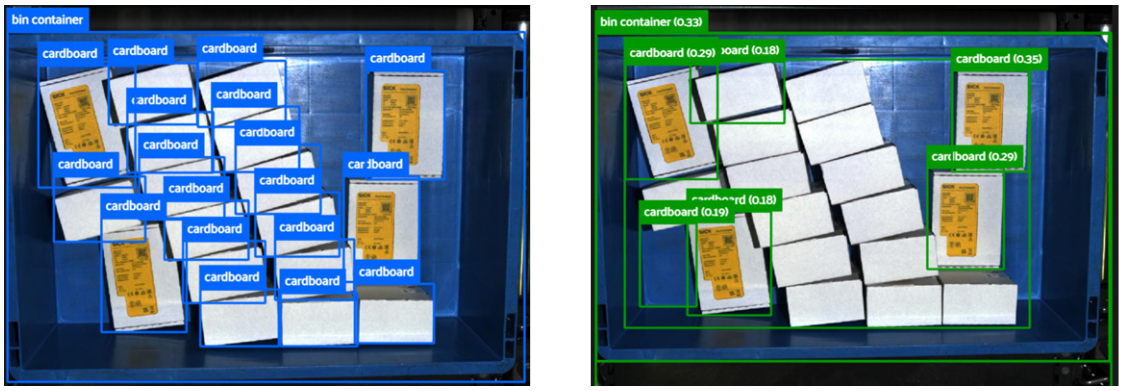


Fig. 3: Ground truth (left) and model predictions of Grounding DINO (right) [5]

References and figures

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations*. OpenReview.net, 2019.

[2] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common Objects in Context. In *Computer Vision - ECCV 2014*, pages 740–755. Springer International Publishing, 2014.

[3] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. In *Computer Vision – ECCV 2024*. Springer Nature Switzerland, 2025.

[4] Matthias Minderer, Alexey A. Gritsenko, and Neil Houlsby. Scaling Open-Vocabulary Object Detection. In *Thirty-seventh Conference on Neural Information Processing Systems*. OpenReview.net, 2023.

[5] Own representation.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010. Curran Associates Inc., 2017.

[7] Ao Wang, Lihao Liu, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. YOLOE: Real-Time Seeing Anything. https://arxiv.org/abs/2503.07465, 2025.

Benutzergeführte Parametrierung von 3rd Party Libraries für eine Web Benutzeroberfläche

Tim Karelin **Jörg Nitzsche**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Robert Bosch Manufacturing Solutions GmbH, Stuttgart-Feuerbach

Einleitung

Der digitale Wandel im Kontext von Industrie 4.0 verändert den Maschinenbau kontinuierlich [5]. Sensorik- und Steuerungskomponenten liefern dabei permanent Maschinendaten. Anwendbar werden diese Daten jedoch erst, wenn sie über Human-Machine-Interfaces (HMIs) verständlich und kontextgerecht visualisiert werden, wie in Abbildung 1 am Beispiel des aktuellen HMI des HMInow-Projekts der Bosch Manufacturing Solutions gezeigt. Gerade im Sondermaschinenbau werden hohe Verfügbarkeits- und Sicherheitsanforderungen gestellt, die auf heterogene Softwaresysteme und knappe Entwicklungsressourcen treffen. Aufgrund dieser Anforderungen lassen sich der Entwicklungsaufwand und das erforderliche Fachwissen nicht immer abdecken. Intuitive, flexibel anpassbare Bedienoberflächen werden somit zum wirtschaftlichen Ziel. Sie beschleunigen Wartungen, senken Servicekosten und steigern die Anlagenverfügbarkeit. Zugleich ermöglicht die Verbindung klassischer Low-Code-Umgebungen für Speicherprogrammierbare Steuerungen (SPS) mit modernen Webtechnologien, dass wissenschaftliche Impulse für moderne Fertigungskonzepte geliefert werden. [6].

Abb. 1: HMInow [1]

Zielsetzung der Arbeit

Ziel dieser Arbeit ist es, einen generischen Visualisierungs-Wizard zu entwerfen und prototypisch umzusetzen, der es Anwendern ohne Webentwicklungs- oder Programmierkenntnisse ermöglicht, vielfältige Visualisierungen in einem HMI, etwa in VisiWin, einer offenen, .NET-basierten HMI/SCADA-Plattform der Inosoft GmbH, zu konfigurieren und zu warten. [4] Vor diesem Hintergrund analysiert die vorliegende Arbeit, wie HMI-Lösungen gestaltet sein müssen, damit wachsende Datenmengen sicher visualisiert werden können und zugleich den steigenden Ansprüchen an Benutzerfreundlichkeit sowie Skalierbarkeit gerecht werden. Die Hauptpunkte sind: 1. Abstrakte Parametrierung: Entwicklung eines schrittweisen Verfahrens zur Parameterspezifikation, das aus den gewählten Einstellungen einsatzbereite Codemodule generiert, ohne Kenntnisse der jeweiligen Bibliotheksyntax vorauszusetzen. 2. Mehrbibliotheksfähigkeit: Bibliotheken werden in einer modularen Architektur gekapselt, so dass sie ohne Änderungen am Wizard leicht ersetzt werden können. 3. Nahtlose Integration in das VisiWin-Hauptprojekt: Es wird revidiert, auf welche Weise die vorproduzierten Module in die VisiWin-Laufzeitumgebung integriert werden, externe Elemente verbinden und im Betrieb dynamische Datenströme aktualisieren können. 4. Corporate-Design-Konformität: Automatische Übernahme des Bosch- und HMI-Designs durch sämtliche generische Ansichten, um eine konsistente Benutzeroberfläche über alle Maschinen hinweg sicherzustellen. Das Ergebnis soll eine funktionsfähige Umsetzung sein, die zeigt, dass klare Abstraktionen den Entwicklungsaufwand reduzieren, Wartungen vereinfachen und die Integration neuer Bibliotheken beschleunigen.

Wizard-MVVM-Architektur

Um die Anforderungen an Flexibilität, Wartung und Skalierbarkeit zu erfüllen, wurde der Wizard unter

Verwendung der Model-View-ViewModel (MVVM) Architektur entwickelt, wie in Abbildung 2 schematisch dargestellt. Dieser Prototyp des Softwaredesigns bietet eine Unterscheidung in Bezug auf die Relevanz von Benutzeroberfläche und Anwendungslogik, sodass das System nachhaltig, modular und erweiterbar ist. Die View des Wizards besteht aus einer statischen Benutzeroberfläche, die die grundlegende Struktur enthält, in die die Views dynamisch geladen werden. Solche User Interface (UI)-Komponenten wie WebViews für die Live-Vorschau und die Fortschrittsanzeige zur Darstellung des aktuellen Fortschritts sind statisch und immer eingebettet und verfügbar, unabhängig vom Navigationsstatus. Jede Views-Seite wird modular realisiert und nur bei Bedarf in das zentrale Fenster integriert. Dieser Ansatz ermöglicht es, Seiten einfach hinzuzufügen, ohne die Kernlogik zu ändern. Die ViewModel-Schicht ist jeweils einem Visualisierungstyp (Diagramm, Tabelle, Liste) zugeordnet. Jede der Darstellungen hat korrespondierende Wizard-Seiten, die in ViewModels gespalten sind. Die Systemtrennung erleichtert die Wartung, da Änderungen oder Erweiterungen präzise und unabhängig von anderen Bereichen vorgenommen werden. Das Model bildet die gesamte Chart-Beschreibungsstruktur ab. Eine zentrale Datei übernimmt die Verwaltung aller Parameter, Eigenschaften und Relationen, die zur Visualisierung einer Grafik erforderlich sind. Die Daten werden über klassische Getter- und Setter-Methoden kontrolliert bereitgestellt. Diese ausbalancierte Struktur ermöglicht es den ViewModels und über die Plugins allen Systemen Flexibilität in der Nutzung und einfaches Handling von Informationen, wenn die Daten überprüft, verarbeitet und trianguliert werden [2].

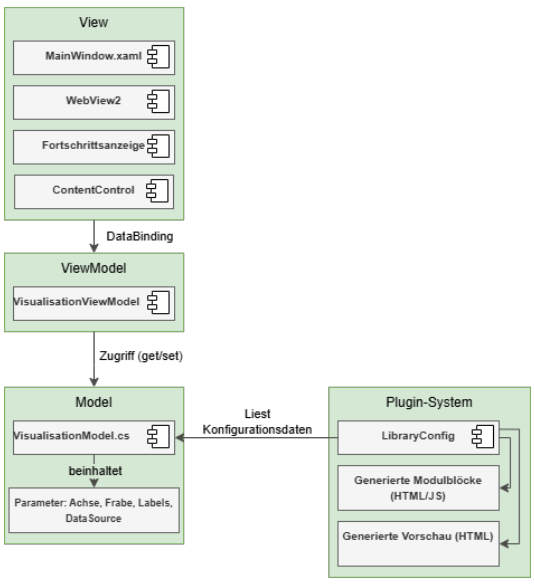


Abb. 2: Diagramm der MVVM-Architektur mit dem Plugin-System [3]

Plugin-System

Die Anbindung von Drittanbieter-Bibliotheken erfolgt über ein eigenständig entwickeltes Plugin-System. Für jede eingebundene Bibliothek wird eine spezifische Konfigurationsklasse definiert, in der festgelegt ist, welche Visualisierungstypen unterstützt werden, welche Parameter auswählbar sind, wie diese in der Benutzeroberfläche dargestellt werden und wie die Live-Vorschau technisch umgesetzt wird. Die Plugins ermöglichen es, neue Bibliotheken flexibel und ohne Anpassung der Kernlogik oder der Wizard-Oberfläche zu integrieren. Zudem generiert das System auf Basis der Plugin-Daten automatisch lauffähige Visualisierungsmodule, die direkt in das Hauptprojekt, insbesondere die VisiWin-Laufzeitumgebung, eingebunden werden können [7].

Umsetzung

Der Wizard wurde mit Windows Presentation Foundation (WPF) und C# unter Verwendung der Model-View-ViewModel (MVVM)-Architektur entwickelt. Das Hauptfenster wurde zuerst als primäre Ansicht erstellt. Es hat einen ContentControl, der Wizard-Seiten dynamisch lädt. Die Navigation durch die Seiten erfolgte mithilfe von Befehlen im jeweiligen MainViewModel. Statische UI-Komponenten wie die WebView2-Instanz für eine Live-Vorschau, eine ProgressBar und Navigationsschaltflächen waren fixiert und in das Layout integriert. Jeder Typ der Visualisierung erhielt seinen eigenen Ordner innerhalb der Projektstruktur. Innerhalb dieser Unterordner befinden sich die entsprechenden Page.xaml-Dateien und ViewModels, die mit INotifyPropertyChanged mit dem zentralen Datenmodell verknüpft sind. Das Model besitzt sämtliche benutzerdefinierte Parameter und dient als Grundlage für die spätere Codegenerierung. Parallel wurde das Plugin-System entwickelt. In einem separaten Plugins-Verzeichnis wurden für jede Third-Party-Bibliothek eigene Konfigurationsklassen implementiert. Diese definieren, welche Parameter unterstützt werden, wie sie im Wizard dargestellt werden und wie die Live-Vorschau zu generieren ist. Die Vorschau wird dabei zur Laufzeit als Hypertext Markup Language (HTML)-Dokument erzeugt und über die WebView2-Komponente eingebunden. Im letzten Schritt wurde die Laufzeitanbindung an VisiWin realisiert. Dazu wurden die generierten Visualisierungsmodule als externe Webkomponenten in das Projekt eingebettet. Über verschiedenste technische Workarounds kann das Modul Maschinendaten aus dem VisiWin-Entwicklungs-Server empfangen und dynamisch visualisieren. So führt die beschriebene Umsetzung zum aktuellen Prototypen, wie er in Abbildung 3 dargestellt ist.

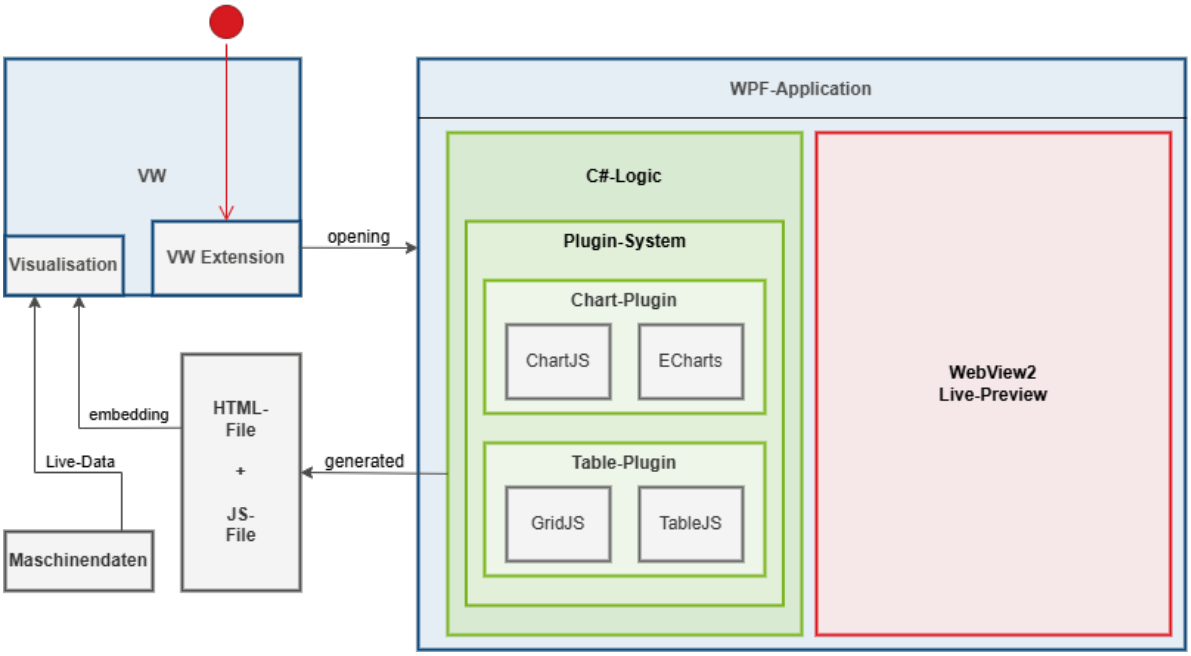


Abb. 3: Aktuelles prototypisches Gesamtsystem des Wizards [3]

Zwischenstand und Ausblick

Der entwickelte Prototyp zeigt deutlich, dass ein generischer Visualisierungs-Wizard erfolgreich implementiert werden kann. Dieser entwickelte Ansatz unterstützt die Einbindung von verschiedenen Third-Party-Bibliotheken zur Erstellung von Visualisierungen aus Nutzereingaben. Diese werden dynamisch in der HMI-Umgebung angezeigt. Zusätzlich wurde eine Schnittstelle zum Laufzeitumfeld VisiWin implementiert, sodass die aktuellen Visualisierungen die zur Laufzeit entstandenen Maschinendaten verarbeiten und aktualisieren können. Trotz der erzielten Ergebnisse sind diese noch nicht als vollständige Automatisierung zu werten. Für jede neue Art gibt es nach wie vor ein spezifisches technisches Gerüst, das manuell erstellt werden muss. Ein nächster Entwicklungsschritt wäre daher, das System so zu erweitern, dass neue Bibliotheken allein über das Einfügen einer Konfigurationsdatei erkannt, geladen und dargestellt werden können. Dies würde die Flexibilität und Skalierbarkeit des Wizards deutlich erhöhen und die Einführung neuer Visualisierungstypen nochmals erheblich vereinfachen.

Literatur und Abbildungen

[1] Robert Bosch Manufacturing Solutions GmbH. HMInow. https://www.bosch-connected-industry.com/de/en/hminow, 2025.

[2] Sudip Chakraborty and Sreeramana Aithal. MVVM Demonstration Using C# WPF. *International Journal of Applied Engineering and Management Letters*, 7:1–14, 2023.

[3] Eigene Darstellung.

[4] Inosoft GmbH. The VisiWin Principle. https://www.inosoft.com/en/product/visiwin-principle/, 02 2025.

[5] Christian Krupitzer, Sebastian Müller, Veronika Lesch, Marwin Zürfle, Janick Edinger, Alexander Lemken, Dominik Schäfer, Samuel Kounev, and Christian Becker. A Survey on Human Machine Interaction in Industry 4.0. https://arxiv.org/abs/2002.01025, 02 2020.

[6] Peter Papcun, Erik Kajati, and Jiri Koziorek. Human Machine Interface in Concept of Industry 4.0. https://ieeexplore.ieee.org/document/8490603, 08 2018.

[7] Jörg Rathlev. Plug-ins: an Architectural Style for Component Software. https://swa.informatik.uni-hamburg.de/files/veroeffentlichungen/Rathlev2008.pdf, 10 2008.

Assoziative Bearbeitungen in der CAD/CAM-Programmierung: Wissensbasierte Automatisierung durch Künstliche Intelligenz

Okan Kizilagil **Jürgen Koch**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma OPUS Entwicklungs und Vertriebs GmbH, Ohmden

Einleitung

In den letzten Jahren hat die Automatisierung industrieller Fertigungsprozesse spürbar an Relevanz gewonnen. Insbesondere durch den zunehmenden Einsatz von Methoden der Künstlichen Intelligenz und das Maschinelle Lernen. Im Bereich der CAM-Programmierung wächst der Bedarf an intelligenten und zugleich anpassungsfähigen Lösungen, die nicht nur effizient arbeiten, sondern auch bestehendes Expertenwissen sinnvoll nutzen. [1] Eine wesentliche Herausforderung dabei ist es, sogenannte Features aus bestehenden Teileprojekten zu erkennen und dessen Bearbeitungsstrategien darauf zu übernehmen. Solche Bearbeitungen, etwa Bohrungen, Taschen oder Schlitze, spiegeln oftmals die gesammelten Erfahrungen und das spezifische Know-how eines CAM-Programmierers wider, das sich über zahlreiche Praxisfälle hinweg aufgebaut hat. [3]
Viele Unternehmen verfügen über große Mengen bereits programmierter CAD-Modelle, die als Teileprojekte archiviert sind. Diese Datenbestände bergen enormes Wissen, das bislang häufig ungenutzt bleibt. Mit Hilfe moderner KI-Technologien bietet sich die Chance, genau dieses Potenzial systematisch zu erschließen. Neue Teileprojekte sollen automatisch auf Basis bestehender Daten programmiert werden. Ein solcher wissensbasierter Ansatz ermöglicht es, Programmierprozesse nicht nur deutlich zu beschleunigen, sondern auch stärker zu vereinheitlichen und qualitativ zu verbessern.

Ziel der Abschlussarbeit

Ziel dieser Abschlussarbeit ist es, das Potenzial zwischen CAM-Programmierung und Künstlicher Intelligenz zu untersuchen und nutzbar zu machen. Hierfür müssen umfangreiche Datenmengen analysiert und strukturiert aufbereitet werden, um sie für das Training von KI-Modellen einsetzen zu können. Das langfristige Ziel besteht darin, vollständige CAD-Modelle automatisiert programmieren zu lassen. Auf diese Weise soll ein intelligenter, KI-gestützter Automatisierungsprozess entstehen, der die Effizienz erhöht, die Wiederverwendbarkeit archivierter Teileprojekte verbessert und die Qualität sowie Konsistenz der Bearbeitungsstrategien in der CAM-Programmierung nachhaltig steigert.
Die innovative Automatisierungstechnik soll direkt in die CAM-Software OPUS eingebunden werden, sodass Assoziative Bearbeitungen automatisch erkannt und umgesetzt werden können. Das Offene Produktions-Unterstützungs-System OPUS ist eine CAD-/CAM-Software zur Erzeugung, Modifikation und Simulation von NC-Programme für Drehmaschinen mit zwei und vier Achsen, Fräsmaschinen mit drei und fünf Achsen, Drahterodier und Brennschneidebearbeitungen im Maschinen-, Werkzeug- und Formenbau. [4]

Vorgehen

Im ersten Schritt wird ein Analysetool entwickelt, das die archivierten Teileprojekte systematisch abarbeitet und auswertet. Dabei werden alle Konturen im CAD-Modell durchlaufen und geprüft, ob ihnen bereits eine Bearbeitung zugeordnet wurde. Die Bearbeitungsinformationen enthalten zentrale Parameter wie zum Beispiel das verwendete Werkzeug, die Drehzahl sowie den Vorschub. Alle erfolgreich analysierten Konturen werden in einer Datenbank abgelegt. Das Analyse Tool wird in der firmeneigenen Programmiersprache SESAM implementiert und bildet die Grundlage für die KI-gestützte Entscheidungsfindung.

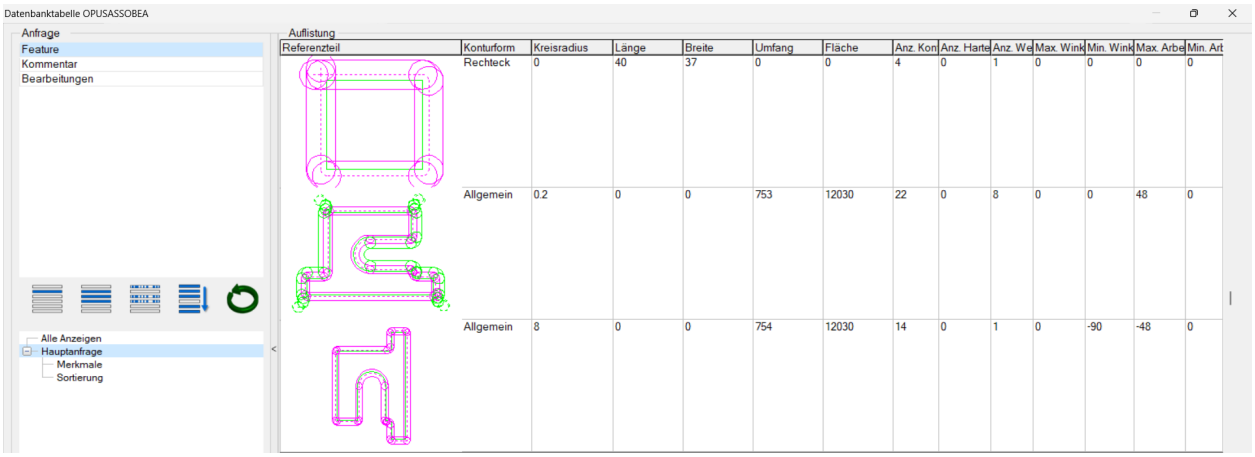


Abb. 1: Datenbanktabelle der Konturdaten [2]

Auf Basis dieser Daten wird ein KI-Modell trainiert, das in der Lage ist zu erkennen und zu entscheiden, wann eine Kontur einer anderen ähnelt. Der Vergleich erfolgt auf Grundlage geometrischer Konturdaten. Wird ein neues CAD-Modell importiert, erfolgt zunächst die automatische Feature-Analyse bei den relevanten Geometrien wie zum Beispiel Bohrungen, Taschen, Schlitze erkannt und an das KI-Modell übergeben werden. Das trainierte KI-Modell prüft ob in früheren Teileprojekten bereits ähnliche Feature bearbeitet wurden. Wenn eine Ähnlichkeit festgestellt wird, gibt das KI-Modell die Referenz der zugehörigen Bearbeitungsstrategie zurück. Zum Abschluss wird die Bearbeitung auf das neue Feature angepasst: Die ermittelte Strategie wird angewendet und die Bearbeitungsbahnen entsprechend der neuen Geometrie automatisch neu berechnet. Folgendes Diagramm stellt die verschiedenen Schnittstellen dar.

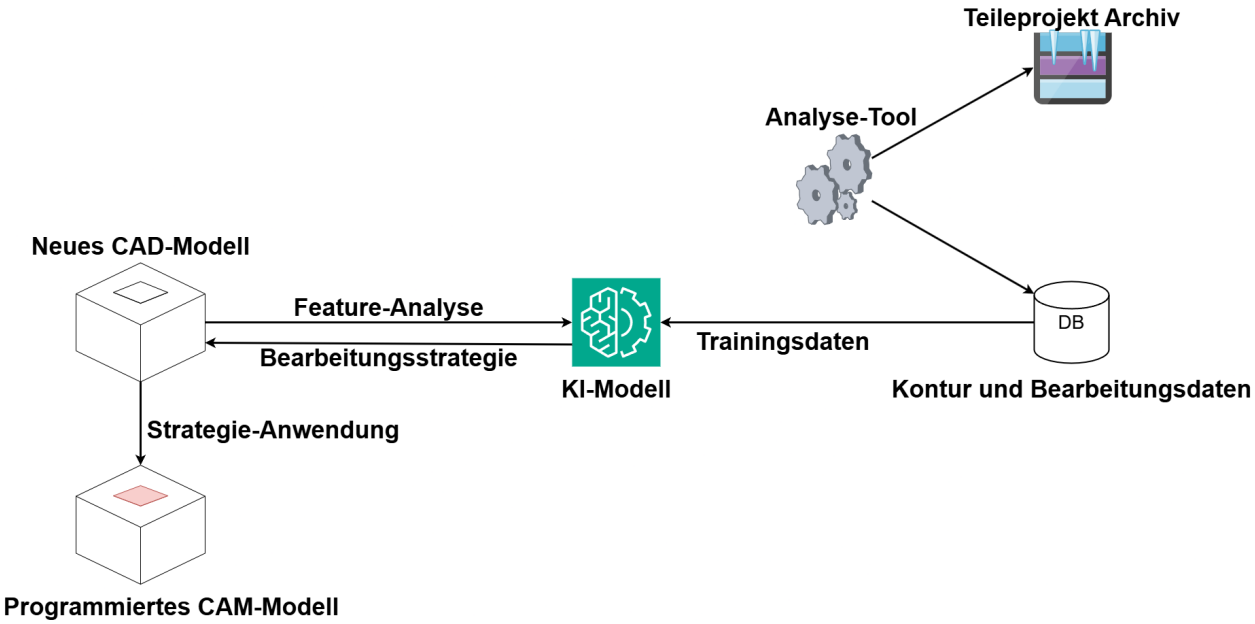


Abb. 2: Zeigt den Prozessablauf zur automatisierten CAM-Programmierung [2]

Vorteile

Durch die Integration künstlicher Intelligenz in der CAM-Programmierung ergeben sich eine Vielzahl an Vorteilen, sowohl auf technischer als auch wirtschaftlicher Ebene. Ein wesentlicher Vorteil liegt in der deutlichen Zeitersparnis bei der Programmierung neuer Teileprojekte. Aufgaben, die bislang manuell durchgeführt wurden, können durch den Einsatz von KI-Modelle strak beschleunigt werden. Darüber hinaus ermöglicht ein solches System eine Standardisierung der Bearbeitungsstrategien. Dadurch kann die Qualität erhöht und verbessert werden. Unterschiede in der

Programmierung einzelner CAM-Spezialisten werden minimiert. Ein weiterer Vorteil ergibt sich aus der Wiederverwendbarkeit archivierter Teileprojekte. Anstatt jedes neue Bauteil von Grund auf zu programmieren, kann auf bestehendes Wissen zurückgegriffen werden. Das wissensbasierte CAM-System kann mit neuen Projekten dazulernen und sich fortlaufend weiterentwickeln.

In der Summe entsteht ein klarer Wettbewerbsvorteil – Unternehmen, die auf KI-gestützte CAM-Systeme setzen, können sich durch Effizienz, Qualität und Innovationsfähigkeit deutlich von der Konkurrenz abheben.

Literatur und Abbildungen

[1] Kristian Arntz, Tobias Claus Brandstätter, Jonas Dorißen, Maik Frye, Jonathan Krauß, Leonie Krebs, Carsten Holst, Rainer Horstkotte, Hendrik Mende, Sven Schiller, Grzegorz Stepien, Moritz Wollbrink, and Zhen Zhen. Künstliche Intelligenz in der Einzel- und Kleinserienfertigung. https://www.ipt.fraunhofer.de/de/publikationen/whitepaper-trendreporte-studien/kuenstliche-intelligenz-in-der-einzel-und-kleinserienfertigung.html, 2021.

[2] Eigene Darstellung.

[3] Willi Gründer and Denis Polyakov. Konstruktionslösungen mit Hilfe von Künstlicher Intelligenz. https://tud.qucosa.de/api/qucosa%3A36932/attachment/ATT-0/, 2019.

[4] DOKUMENTATION OPUS. OPUS (CAM-Software). https://de.wikipedia.org/wiki/OPUS_(CAM-Software), 2023.

Entwicklung eines Tools für die Variantenverwaltung und automatische Konfigurierung von AUTOSAR Software mit Vector Davinci Configurator

Marc Klein **Michael Scharf**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma SEG Automotive Germany GmbH, Stuttgart

Einleitung

Moderne Autos sind komplexe Maschinen mit hunderten Steuergeräten, die nach Herstelleranforderung von verschiedenen Zulieferern entwickelt werden. Wenn für die Realisierung einer Fahrzeugfunktion (z.B. Aktivierung der Innenbeleuchtung nach Deaktivierung der Zündung) mehrere Steuergeräte von verschiedenen Zulieferern zusammenarbeiten müssen, erforderte dies in der Vergangenheit einen großen Koordinierungsaufwand und umfangreiche Integrationstests. [4]
Um diese Probleme zu vereinfachen, wurde 2003 von führenden Automobilherstellern und Zulieferern der AUTOSAR Standard (**AUT**omotive **O**pen **S**ystem **AR**chitecture) geschaffen.

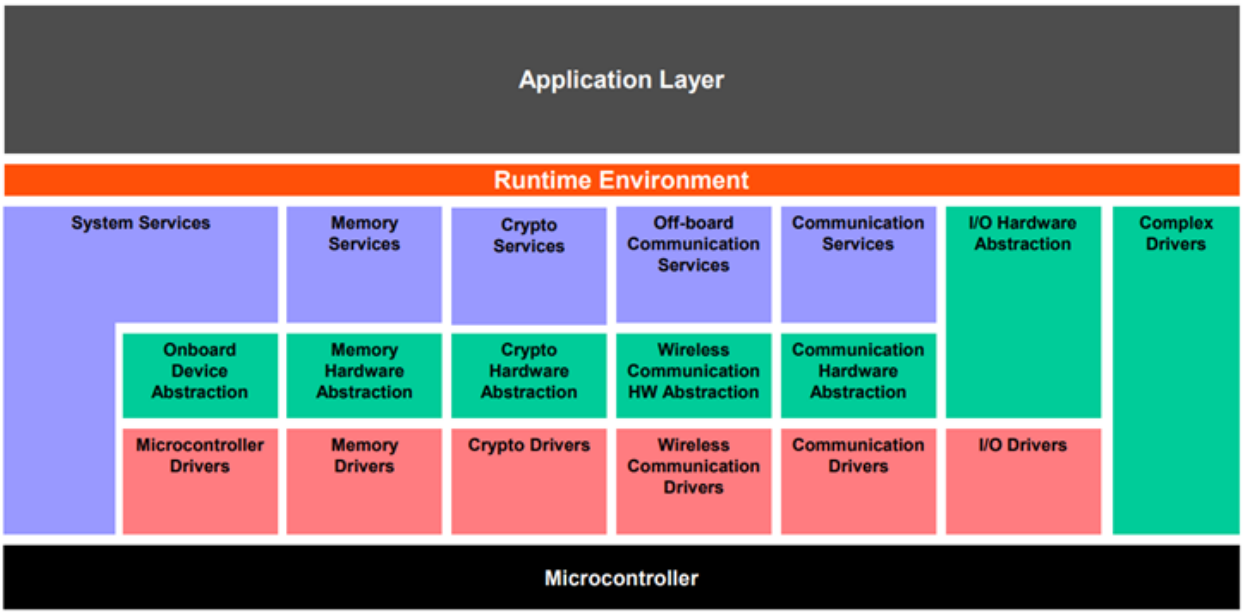


Abb. 1: AUTOSAR Softwarearchitektur [2]

Zusätzlich zu einer standardisierten Software-Architektur führte AUTOSAR eine einheitliche Methodik für die Entwicklung von Automobilsoftware ein. Alle Aspekte einer Softwarekomponente werden im AUTOSAR Schema (.arxml) beschrieben. Die tatsächlichen Softwarekomponenten werden aus diesen Funktionsbeschreibungen fast vollständig automatisch generiert. Damit wird der Entwicklungs- und Testaufwand erheblich reduziert sowie die Codequalität und Konformität mit Softwarequalitätsstandards sichergestellt. [3]

Motivation

Die Steuergerätesoftware wird in einem iterativen Prozess geschaffen, weiterentwickelt und optimiert.

Auf Basis einer Grundversion werden kundenspezifische Softwarevarianten entwickelt, um individuellen Spezifikationen gerecht zu werden. Im Umfang dieser Arbeit soll ein Tool entwickelt werden, um die Verwaltung dieser Varianten zu vereinfachen. Optimierungen, die bei der Entwicklung von Kundenvarianten entstehen, sollen möglichst einfach in die Grundversion übernommen und weiter in alle anderen Softwarevarianten verteilt werden können. Die auf XML-basierenden AUTOSAR Schema Dateien sind durch die große Anzahl von Elementabhängigkeiten praktisch nicht direkt vergleichbar, wodurch das Nachvollziehen von Versionsunterschieden nicht direkt möglich ist.

Vorgehensweise

SEG-Automotive verwendet Vector DaVinci Configurator zur Konfigurierung und Generierung der Steuergerätesoftware aus den Funktionsbeschreibungen.

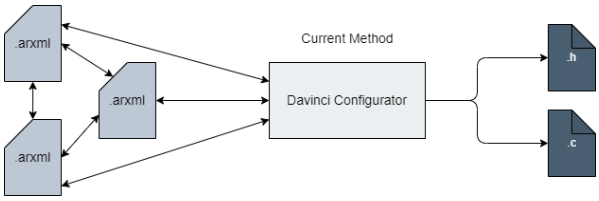


Abb. 2: Codegenerierung aus Funktionsbeschreibung [1]

Da die Unterschiede zwischen den Varianten aufgrund ihrer Komplexität nicht direkt in ihren Funktionsbeschreibungen erkennbar sind, muss eine dem Modell entsprechende Zwischenvariante erstellt werden.
DaVinci Configurator verfügt über umfangreiche APIs im Rahmen des DaVinci Configurator „AutomationInterface“. Außerdem erlaubt DaVinci Configurator die Ausführung von Groovy oder Java Nutzerskripten in seiner internen „Skript Engine“. Teil des „AutomationInterface“ Pakets sind eine Reihe von Funktionsbibliotheken für die genannten Sprachen, die eine direkte Verwendung der APIs innerhalb der Skripte ermöglichen. Damit ermöglicht das „AutomationInterface“ für DaVinci Configurator die Automatisierung vieler Arbeitsschritte und die Implementierung dieser Arbeit.

Für die Implementierung dieser Arbeit wird hauptsächlich die „Model API“ verwendet, um die Modelldetails in einem JSON-Format zu exportieren und einen elementweisen Vergleich zu ermöglichen. Für den Export werden die gewählten Elemente oder Module über ihr „Bswmd Model“ aufgerufen, ihre Parameter, Werte und Referenzen ausgelesen und elementweise in einer JSON-Datei gespeichert. Zwischengespeicherte JSON-Dateien können dann unkompliziert Element für Element verglichen werden. Die beim Vergleich der JSON-Dateien entstehende zusammengeführte Variante soll anschließend wieder in ein vollständiges DavVinci Configurator Modell umgewandelt werden.

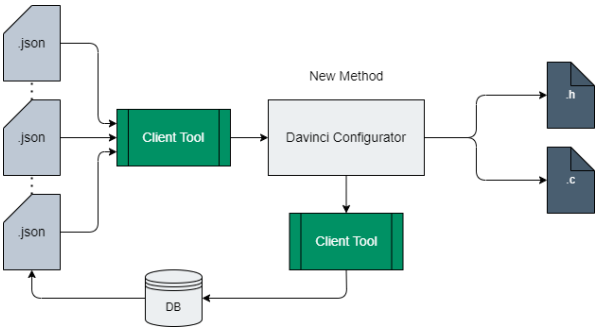


Abb. 3: Codegenerierung aus JSON-Dateien [1]

Dies kann ebenfalls mit der „Model API“ realisiert werden. Über die „Transactions API“ können Elemente erstellt und modifiziert werden. Ein zwischengespeichertes vollständiges Modell könnte so als Vorlage für zukünftige Projekte eingesetzt werden und leicht auf dem neuesten Stand der Entwicklung gehalten werden.

Ausblick

Zusätzliches Ziel der Arbeit ist es, die Übernahme simpler Änderungen zu automatisieren. Sollte sich eine Referenz einer Komponente ändern, so soll dies automatisch auf andere relevante Komponenten übernommen werden können.
Langfristig soll das Tool die Arbeit der Softwareentwicklung und Versionsverwaltung durch eine Simplifizierung der Abläufe und der Automatisierung von Routineaufgaben unterstützen und vereinfachen.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Autosar Konsortium. Architecture - Overview of Software Layers. https://www.autosar.org/fileadmin/standards/R24-11/CP/AUTOSAR_CP_EXP_LayeredSoftwareArchitecture.pdf, 11 2024.

[3] Autosar Konsortium. Autosar classic platform. https://www.autosar.org/standards/classic-platform, 2025.

[4] Helmut Schelling. AUTOSAR - Für Alles Gewappnet. https://cdn.vector.com/cms/content/know-how/_technical-articles/AUTOSAR/AUTOSAR_SG_Mobility2.0_201402_PressArticle_DE.pdf, 2014.

Remote-Entwicklung in Containern: Effizientes Entwickeln und Debuggen von Microservices

Moritz Kuebler **Dennis Grewe**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Cenit AG, Stuttgart

Einleitung und Problemstellung

In den letzten Jahren hat sich die Entwicklung und Bereitstellung von Webanwendungen grundlegend verändert. Zunehmende Anforderungen an Skalierbarkeit und Flexibilität der Anwendungen und den Wunsch nach häufigeren und zeitnahen Updates und Erweiterungen, sorgen dafür, dass moderne Webanwendungen vermehrt auf einer Microservice-Architektur basieren. Im Vergleich zur herkömmlichen monolithischen Architektur, bei der eine Anwendung als komplette Einheit erstellt wird, beruht das Prinzip der Microservices darin, die gesamte Anwendung in einzelne, spezialisierte Dienste aufzuteilen, welche isoliert voneinander entwickelt, getestet und bereitgestellt werden können (siehe Abbildung 1). [4].

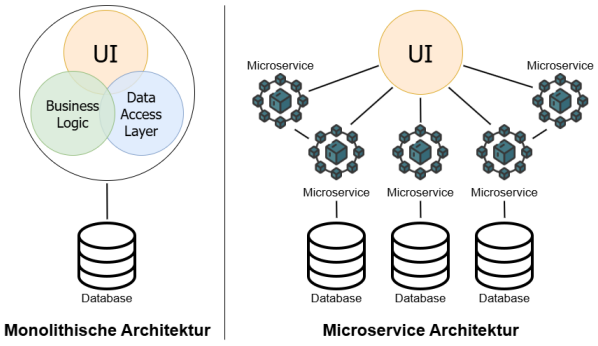


Abb. 1: Monolithischen Anwendung vs. Microservice-Architektur [3]

Für die Bereitstellung dieser Microservices haben sich Technologien wie Docker und Kubernetes inzwischen etabliert. Im Gegensatz zur traditionellen Bereitstellung mittels physischer oder virtueller Maschinen, bei der ein komplettes Betriebssystem zur Ausführung der Anwendung eingerichtet und betrieben werden muss, ermöglicht Docker das Verpacken der einzelnen Dienste in sogenannten Containern, welche den ausführbaren Dienst inklusive aller benötigten Abhängigkeiten beinhalten. Die Container können dadurch, nach einmaligem Erstellen, flexibel auf verschiedenen Systemen verwendet werden, ohne dass eine manuelle Konfiguration benötigt wird [7].
Mit Kubernetes lassen sich Container zusätzlich effizient und einfach verwalten. Dafür wird ein Cluster, also ein Zusammenschluss aus mehreren physischen oder virtuellen Maschinen, erstellt, in welchem die Container ausgeführt werden. Diese können dadurch auf mehrere Maschinen verteilt und automatisch hoch- und runtergefahren werden, wodurch Dienste skaliert werden können, um beispielsweise den Ausfall eines stark ausgelasteten Dienstes zu verhindern. Kubernetes erleichtert außerdem die Kommunikation und Interaktion zwischen mehreren Containern, wodurch die einzelnen Microservices zu einer vollständigen Anwendung verwachsen [6].
Auch beim Aufsetzen einer Entwicklungs- oder Testumgebung auf den lokalen Geräten der Entwickler können diese Technologien eingesetzt werden. In der Praxis wird während des Entwicklungsprozesses jedoch in vielen Fällen weiterhin auf eine lokale Einrichtung und ein manuelles Starten der einzelnen Microservices gesetzt. Das liegt unter anderem daran, dass bei vielen Entwicklern das Wissen und die Erfahrung im Umgang mit Docker und Kubernetes fehlt. Dies führt zu einem hohen manuellen Aufwand für den Entwickler und kann zu Problemen durch Unterschiede zwischen der lokalen Entwicklungs- und der Container-basierten Produktivumgebung führen. Da die lokalen Rechner der Entwickler manuell eingerichtet werden, kann nicht sichergestellt werden, dass jeder Entwickler die gleichen Tools in denselben Versionen verwendet, wie die Container. Außerdem werden die Microservices während der Entwicklung nicht isoliert voneinander, sondern in einer einheitlichen Umgebung ausgeführt, was beispielsweise dazu sorgt, dass alle Microservices auf dieselben Ressourcen, wie Verzeichnisse oder andere Microservices, zugreifen können, während dies in einer containerisierten Umgebung korrekt konfiguriert sein muss. Das kann dazu führen, dass Funktionen auf den Entwicklerrechnern problemlos funktionieren, jedoch

nach der Bereitstellung in der Produktivumgebung Fehler verursachen.

Zielsetzung

Diese Bachelorarbeit untersucht die Möglichkeit den gesamten Entwicklungsprozess einer modernen Webanwendung zu optimieren, indem die Microservices zum Testen und Debuggen nicht auf den lokalen Rechnern ausgeführt werden, sondern in einer Entwicklungsumgebung, welche so weit wie möglich der Produktivumgebung der Anwendung entspricht. Das bedeutet, die Microservices sollen auch während der Entwicklung in isolierten Containern ausgeführt werden. Die Container sollen dabei nicht direkt auf den Entwicklerrechnern laufen und verwaltet werden, sondern in einer externen Umgebung, beispielsweise einem Kubernetes Cluster. Dabei soll es keine Rolle spielen, ob es sich um eine Cloud Umgebung von Amazon, Google, usw. handelt oder ob diese selbst verwaltet wird. Der Entwickler muss seinen Code mit dieser remote Umgebung synchronisieren und zum Testen und Debuggen darauf zugreifen können. Um dies zu erreichen, sollen mehrere Tools untersucht werden und auf Grundlage dieser Recherche ein Konzept erarbeitet und bereitgestellt werden, mit Hilfe dessen eine Remote-Entwicklungsumgebung bereitgestellt werden kann. Die Erstellung, Konfiguration und Verwendung dieser soll für den Entwickler so einfach wie möglich gestaltet werden, zum Beispiel mit Hilfe einer simplen Konsolenanwendung. Die Ergebnisse der Arbeit sollen das Unternehmen Cenit AG dabei unterstützen, den Prozess bei der Entwicklung von Webanwendungen zu vereinfachen und beschleunigen.

Vorgehensweise

Zu Beginn wird für die Analyse des Ist-Zustands und zur Ermittlung von genauen Anforderungen ein Fragebogen an Entwickler und Verantwortliche der Cenit AG gesendet, die aktuell oder in Zukunft im Entwicklungsprozess der Webanwendung beteiligt sind. Ziel ist es, die bestehenden Arbeitsabläufe zu analysieren, aktuelle Herausforderungen zu identifizieren und die Erwartungen an einen verbesserten Entwicklungsprozess zu erfassen, um daraus konkrete Anforderungen zu definieren. Anschließend erfolgt auf Grundlage der gesammelten Anforderungen und Vorgaben des Unternehmens eine Recherche über bestehende Technologien und Tools, die eine Remote-Entwicklung ermöglichen [5]. Die Ergebnisse der Recherche dienen dazu zwei Ansätze zu entwerfen, wie der Entwicklungsprozess der Webanwendung angepasst werden könnte.

Ansatz 1: Interaktive Entwicklungsumgebung im Container

Die untersuchten Tools „DevSpace“ und „Okteto“ ermöglichen die Bereitstellung von Containern, die eine lokale Entwicklungsumgebung simulieren, jedoch vollständig in einem Kubernetes Cluster laufen und mit den darin laufenden Microservices kommunizieren können. Der Entwickler verbindet sich über ein Terminal oder eine integrierte Entwicklungsumgebung (z. B. VS Code Remote) mit diesen Containern und kann darin Änderungen vornehmen, Befehle ausführen und Anwendungen starten oder debuggen [2]. Dieser Ansatz bietet dem Entwickler viel Kontrolle und Flexibilität, erfordert jedoch ein gewisses Maß an Wissen im Umgang mit Containern und Linux-Umgebungen, da Scripte oder manuelle Befehle nötig sind, um den Entwicklungscontainer zu konfigurieren und den Microservice darin zu starten. Für die Umsetzung dieses Ansatzes hat sich DevSpace als das geeignetere Tools gegenüber Okteto erwiesen. Bei Okteto lassen sich Container, an denen nicht entwickelt wird, welche aber zum Testen und Debuggen anderer Dienste benötigt werden, nicht automatisch über Okteto auf dem Cluster ausführen, wenn die Okteto Plattform im verwendeten Kubernetes Cluster nicht installiert ist, was nicht gewünscht ist, da es extra Aufwand und einen höheren Ressourcenverbrauch im Cluster bedeuten würde. DevSpace bietet diese Funktion, ohne, dass das Kubernetes Cluster dafür speziell eingerichtet werden muss.

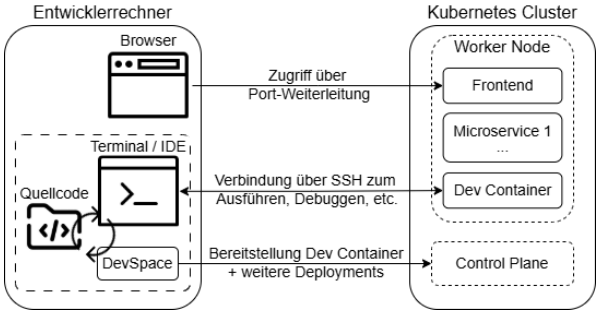


Abb. 2: Darstellung Ansatz 1 mit DevSpace [3]

Ansatz 2: Entwicklung durch kontinuierliche Container-Bereitstellung

Der zweite Ansatz basiert auf den Tools „Tilt“ oder „Skaffold“. Wie beim ersten Ansatz können Container mit Hilfe dieser Tools schnell und unkompliziert in einem Kubernetes Cluster bereitgestellt werden und lokale Dateien in Echtzeit in den Containern synchronisiert werden. Bei diesem Konzept dienen die Container jedoch nicht als interaktive Entwicklungsumgebung, sondern lediglich zur Ausführung der Microservices.

Änderungen sorgen dafür, dass die Container bei Bedarf automatisch neu erstellt und bereitgestellt werden oder die Änderungen zur Laufzeit in den Container kopiert werden [8] [1]. Der Entwickler verbindet sich dabei nicht direkt mit den Containern und führt keine manuellen Befehle oder Scripte darin aus. Die Container werden im Vorfeld über Dockerfiles konfiguriert und führen die Microservices beim Start automatisch aus. Im Vergleich zum ersten Ansatz ist hierbei weniger Wissen der einzelnen Entwickler im Umgang mit Containern und Linux erforderlich. Es muss aber darauf geachtet werden, dass die Dockerfiles zum Erstellen der Container korrekt eingerichtet sind, um beispielsweise das Debuggen der darin laufenden Microservices zu ermöglichen. Sowohl Tilt als auch Skaffold eignen sich gut, um diesen Ansatz umzusetzen. Während die Konfiguration mit Skaffold etwas übersichtlicher und einfacher ist, bietet Tilt ein paar weitere Funktionen, wie eine Benutzeroberfläche, in der direkt auf die Logs aller laufenden Container zugegriffen werden kann und die Möglichkeit mehrere Port-Weiterleitungen für einen Container direkt in der Konfigurationsdatei einzurichten, was beispielsweise für das Remote-Debugging benötigt wird. Daher wird Tilt für die Umsetzung dieses Ansatzes verwendet.

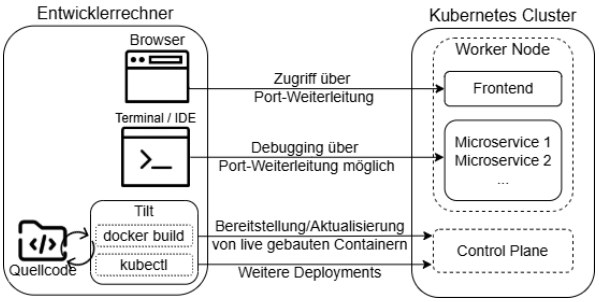


Abb. 3: Darstellung Ansatz 2 mit Tilt [3]

Ausblick

Beide Ansätze sollen mit Hilfe der Tools „DevSpace“ bzw. „Tilt“ prototypisch umgesetzt und anschließend von einzelnen Entwicklern getestet werden. Die Entwickler bekommen dazu eine kurze Einweisung und testen die Funktionsfähigkeit und Handhabung der Prototypen durch beispielhafte Code-Änderungen an einem ausgewählten Microservice. Mittels eines Fragebogens sollen die Entwickler Feedback zu den beiden Ansätzen geben. Anschließend soll auf Basis des Feedbacks eine Entscheidung getroffen werden, welches Konzept in Zukunft für die Entwicklung verwendet werden soll. Bevor der Entwicklungsprozess aller Entwickler auf dieses Konzept umgestellt wird, soll eine Möglichkeit, in Form einer simplen Konsolenanwendung, geschaffen werden, womit die Entwickler das Tool und die dafür benötigten Konfigurationsdateien für ihre Anforderungen einrichten können.

Literatur und Abbildungen

[1] Phil Adam. Die Bedeutung von Optimierungswerkzeugen in der Softwareentwicklung mit Kubernetes und deren Auswirkung auf den Entwicklungsprozess, 2020.

[2] Sayanta Banerjee. Five tools to increase Kubernetes developer productivity. https://www.civo.com/blog/five-tools-to-increase-kubernetes-developer-productivity, 02 2022.

[3] Eigene Darstellung.

[4] Hanno Kortekamp. Vor- und Nachteile von monolithischen und Microservice-Architekturen. https://www.arvato-systems.de/blog/microservices-vs-monolith, 10 2024.

[5] Stefan Mückstein et al. Kubernetes Tools: Minikube, kind, skaffold, tilt, devspace und deren Platz im Entwicklungsprozess. https://cloudomation.com/de/blog/kubernetes-tools-uebersicht/, 09 2024.

[6] Annika Opitz. Was ist Kubernetes? https://www.plusserver.com/blog/kubernetes/, 11 2022.

[7] Stephanie Susnjara and Ian Smalley. Was sind Container? https://www.ibm.com/de-de/topics/containers, 05 2024.

[8] Andreas Zitzelsberger. Turnaround-Turbo: Effizientes Shift-Left DevSecOps mit Java und Tilt. https://www.sigs.de/artikel/turnaround-turbo-effizientes-shift-left-devsecops-mit-java-und-tilt/, 01 2025.

UX-Optimierung eines Wissensmanagement-Portals

Robert Lang **Astrid Beck**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Im Rahmen einer Bachelorarbeit wird derzeit die Optimierung eines konzerninternen Wissensmanagementsystems untersucht, das als zentrale Plattform für das Wissensmanagement dient. Im Fokus steht ein nutzerzentrierter, iterativer Redesign-Prozess dessen erste Phase, eine heuristische Evaluation, aktuell durchgeführt wird. Aufbauend auf dieser Analyse sollen später Nutzungsumfragen und A/B-Tests mit interaktiven Prototypen folgen. In diesem Artikel wird beschrieben, wie der Untersuchungsrahmen konzipiert ist, welche theoretischen Grundlagen herangezogen werden und welche Ziele durch ein überarbeitetes Overlay erreicht werden sollen. Es wird erwartet, dass der finale Prototyp die Effizienz der Wissensnutzung signifikant steigern und die Benutzerzufriedenheit messbar verbessern wird.

1 Einleitung

In modernen Unternehmen wird digitalen Wissensplattformen eine essenzielle Rolle zugeschrieben, wenn es darum geht, Informationen zentral zu sammeln, zu strukturieren und bereitzustellen. Allerdings bleibt dieses Potenzial häufig ungenutzt, da viele Portale durch eine übermäßige Komplexität, unklare Navigation und mangelnde Rückmeldungen auffallen. Auch im vorliegenden Fall, dem Wissensmanagementsystems eines Großunternehmens, lassen sich durch erste Beobachtungen, Defizite in der Benutzerfreundlichkeit feststellen. Die Menüstruktur erscheint überfrachtet, Begriffe werden inkonsistent verwendet, und Rückmeldungen auf Systemaktionen bleiben aus.
Durch einen systematisch angelegten, nutzerzentrierten Designprozess soll ein Overlay entwickelt werden, das die alltägliche Nutzung erleichtert und sowohl funktional als auch visuell optimiert ist. Anders als in rein technischen Dokumentationen wird in dieser Arbeit ein besonderes Augenmerk auf die Nachvollziehbarkeit der Herangehensweise gelegt – sowohl aus theoretischer Sicht als auch im Hinblick auf gestalterische Entscheidungen.

2 Theoretischer Rahmen und Zielsetzung

Die methodische Grundlage der Arbeit bildet das Heuristik-Modell von Jakob Nielsen [2], dessen zehn Usability-Prinzipien, darunter Konsistenz, Sichtbarkeit des Systemstatus und Fehlervermeidung, aktuell zur Bewertung des bestehenden Portals herangezogen werden. Ergänzt wird dieser Ansatz durch die Cognitive Load Theory von John Sweller [4], der zufolge unnötig komplexe Informationsstrukturen das Arbeitsgedächtnis belasten und die Effizienz der Nutzung beeinträchtigen. Zusätzlich wird auf Donald A. Norman [3] verwiesen, der betont, dass mentale Modelle mit dem Systembild übereinstimmen müssen, um Fehlbedienungen zu vermeiden.
Ziel ist die Gestaltung eines konsistenten Overlays, das kognitive Belastung reduziert und Orientierung verbessert. Langfristig soll ein UI-/UX-Manifest entstehen, das Gestaltungsrichtlinien dokumentiert und laufend weiterentwickelt wird.

3 Geplanter methodischer Ablauf

Die Umsetzung wird in mehreren Phasen strukturiert erfolgen. In der ersten Phase, der heuristischen Evaluation, wird das bestehende Portal anhand der Usability-Heuristiken analysiert. Darauf aufbauend soll eine Nutzungsumfrage entwickelt und verteilt werden, die qualitative und quantitative Rückmeldungen zur Nutzung, zu Erwartungen und zu subjektiven Eindrücken erfassen soll. Im Anschluss ist vorgesehen, zwei interaktive Prototypen mit Figma zu erstellen, welche von Nutzenden im Hinblick auf Bedienbarkeit und Ästhetik getestet werden sollen. In einer vierten Phase wird ein direkter Vergleich dieser Prototypen (A/B-Test) durchgeführt, bei dem Kennzahlen wie Task Completion Rate und Bearbeitungsdauer ermittelt werden. Abschließend sollen alle Erkenntnisse systematisch in einem UI-/UX-Manifest dokumentiert werden. Der gesamte Prozess wird iterativ angelegt, sodass auf Zwischenergebnisse flexibel reagiert und Gestaltungsentscheidungen angepasst werden können.

4 Vorläufige Beobachtungen aus der heuristischen Analyse

Bereits im Rahmen der laufenden heuristischen Evaluation lassen sich erste Schwächen identifizieren. So scheint die Navigation des Portals nicht nur überladen zu sein, sondern widerspricht auch etablierten Navigationsmustern. Die Bezeichnungen einzelner Menüpunkte erscheinen je nach Kontext variierend, was zu Verunsicherung führt. Rückmeldungen auf Nutzerinteraktionen, etwa beim Absenden von Formularen, fehlen vollständig. Während die Nutzungsumfrage noch vorbereitet wird, sollen bereits erste Figma-Prototypen zur Visualisierung möglicher Alternativen beitragen. Dabei wird ein Ansatz verfolgt, bei dem eine Variante auf minimalistisches Design setzt, während die andere auf zusätzliche kontextsensitive Hilfestellungen fokussiert.

5 Analyse des Moduls Wissensmanagement“

Im Modul „Wissensmanagement“ wurde beobachtet, dass sich Navigationselemente nur schwach vom Hintergrund abheben, wodurch die visuelle Orientierung erschwert wird. Zudem stellt eine tief verschachtelte Menüstruktur mit über fünf Ebenen eine erhebliche Hürde für effizientes Arbeiten dar. Exemplarische Navigationspfade zeigen, wie sich Inhalte wie „Level 1“, „Level 2“, „Level 3“, „Level 4“ und „Level 5“ über mehrere Hierarchieebenen erstrecken. In einer späteren Version des Overlays wird angestrebt, durch eine reduzierte Menüstruktur und eine optimierte Farbgebung die Lesbarkeit und Orientierung spürbar zu verbessern.

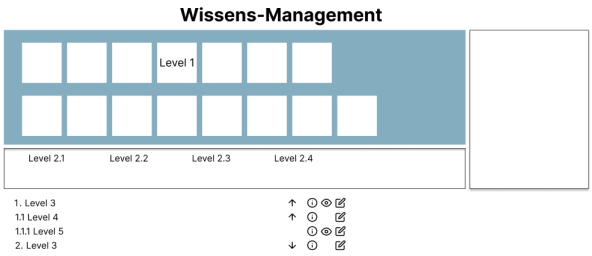


Abb. 1: Figma-Abbildung der Wissensmanagement-Seite: geringe Kontrastierung und unübersichtliche Menühierarchie [1]

6 Erste Erkenntnisse aus dem Modul „Lessons Learned“

Auf der Lessons-Learned-Seite wird ein weiteres Problem sichtbar: Eine Vielzahl an Filteroptionen erschwert die gezielte Informationsauswahl. Zusätzlich wird ein Wechsel zwischen deutscher und englischer Sprache festgestellt, der sich störend auf den Lesefluss auswirkt. Auch die Tabellenstruktur erscheint übermäßig komplex. Das Aufrufen eines Beitrages innerhalb der Tabelle ist derzeit ausschließlich über eine bestimmte Spalte möglich, was eine zusätzliche kognitive Hürde darstellt. In der künftigen Überarbeitung wird eine sprachliche Vereinheitlichung sowie eine übersichtlichere Darstellung der Filterelemente angestrebt.

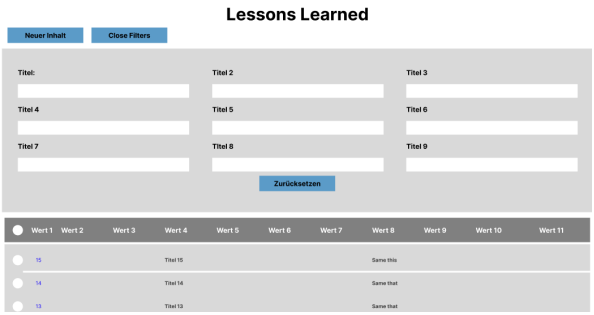


Abb. 2: Vereinfachte Figma-Abbildung der Lessons Learned-Seite: überladene Filter, uneinheitliche Sprachmischung und komplexe Tabellenstruktur. [1]

7 Entwurf des zukünftigen Optimierungskonzepts

Aus der Synthese der vorangegangenen Prototypen wird ein hybrides Overlay entstehen, das die jeweiligen Stärken kombiniert. In diesem Prototyp C sollen maximal drei Menüebenen eingeführt und alle Kategorien klar bezeichnet werden. Nutzeraktionen werden durch visuelle Rückmeldungen – beispielsweise Ladeindikatoren und kurze Bestätigungen – begleitet. Die Terminologie wird konsistent anhand des unternehmensinternen Glossars gestaltet. Für erfahrene Nutzerinnen und Nutzer wird geplant, kontextsensitive Tooltips sowie Tastenkombinationen zu integrieren, um effizientere Workflows zu ermöglichen.

8 Erwartete Effekte und weitere Schritte

Es wird erwartet, dass der finale Prototyp die durchschnittliche Bearbeitungszeit um mindestens 25 % reduzieren wird. Gleichzeitig soll ein System Usability Scale (SUS)-Wert von über 80 erreicht werden. In zukünftigen Arbeitsschritten werden Optimierungen für mobile Endgeräte sowie die Integration von Touch-Gesten berücksichtigt. Darüber hinaus ist geplant, eine semantische Suchfunktion einzuführen, die eine kontextbasierte Navigation erleichtert. Zur Sicherstellung einer kontinuierlichen Verbesserung wird ein regelmäßiger Evaluationszyklus etabliert werden, in

dem das Manifest an neue Anforderungen angepasst werden kann.

9 Schlussfolgerung

Die in dieser Arbeit verfolgte Herangehensweise zeigt auf, wie ein nutzerzentrierter, iterativer Redesign-Prozess strukturiert geplant und durchgeführt werden kann. Während sich das Projekt noch in einer frühen Phase befindet, lassen sich bereits grundlegende Schwächen des bestehenden Systems identifizieren und erste Gestaltungsprinzipien formulieren. Das geplante Hybrid-Overlay verbindet funktionale Effizienz mit visueller Klarheit und legt damit die Grundlage für eine nachhaltige UX-Strategie. Durch die Etablierung eines kontinuierlichen Evaluationsprozesses wird sichergestellt, dass das Wissensmanagementsystem langfristig den sich wandelnden Anforderungen gerecht werden kann.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Jakob Nielsen. *Usability Engineering*. Academic Press, 1993.

[3] Donald Norman. *The Psychology of Everyday Things*. Basic Books, 1988.

[4] John Sweller. Cognitive load during problem solving: Effects on learning. *Taylor & Francis im Auftrag der Cognitive Science Society*, pages 257–285, 1988.

Code-Metriken-gestützte Restrukturierung und Modularisierung einer großen und gewachsenen C++-Codebasis

Noel Leyrer **Jörg Nitzsche**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma TeamViewer Germany GmbH, Göppingen

Einleitung

Wenn eine Software im Laufe der Jahre immer mehr Features erhält, wächst auch der Code. Am Anfang ist alles noch überschaubar, aber die Komplexität kann sehr schnell zunehmen. So gibt es oft Abhängigkeiten (Dependencies) zwischen Features, die eigentlich unabhängig voneinander funktionieren sollten. Das Ziel dieser Arbeit ist es, diese Abhängigkeiten in einer großen C++-Codebasis sichtbar zu machen und zu entfernen. Dadurch soll eine bessere Modularisierung erreicht werden.

Motivation und Zielsetzung

Es gibt viele Gründe für Abhängigkeiten, z.B. das Streben nach schneller und effizienter Entwicklung, ggf. mit Workarounds [4]. Das kann funktionieren, solange nichts am bestehenden Code geändert werden muss. Genau diese Voraussetzung ist aber in der Praxis nicht gegeben, da auch bestehende Features wieder geändert werden müssen. Hier entsteht dann sehr schnell das Problem, dass man bei einer kleinen Änderung nicht mehr absehen kann, was diese auslöst. Es treten dann Fehler an Stellen auf, die eigentlich nichts damit zu tun haben sollten. Zu diesem Zeitpunkt ist es bereits zu spät, diese sogenannten Technical Debts zu beheben [4]. Im Application Feature Layer des C++-Projekts soll eine Modularisierung durch Neustrukturierung aller Projekte erreicht werden. Begleitend soll der Fortschritt durch geeignete Code-Metriken gemessen werden. Am Ende soll folgende Forschungsfrage beantwortet werden: *Wie kann die Wirksamkeit der Umstrukturierung und Modularisierung einer bestehenden Codebasis mit Hilfe von Code-Metriken gemessen werden?*

Buildsystem und Projektstruktur

Die erwähnte C++-Codebasis umfasst Code für verschiedene Plattformen wie Windows, Linux, macOS, iOS und Android. Durch die Verwendung des Tools *CMake* kann für jede Plattform passend gebaut werden. Das Tool ist ein Meta-Buildtool und kann Buildkonfigurationen für verschiedene Buildtools oder IDEs wie Microsoft Visual Studio oder Apple Xcode erstellen. Mit diesen Konfigurationen kann der Code dann für verschiedene Zielplattformen mit Compilern wie MSVC und Clang übersetzt werden. So können verschiedene Plattformen oder Toolchains (z.B. x86, ARM) unterstützt werden. Die Konfigurationen werden in *CMakeLists.txt* angelegt und beschreiben die Targets, Abhängigkeiten und Buildoptionen. Sie werden hierarchisch in jedem Projekt erstellt und aufgerufen. [2]

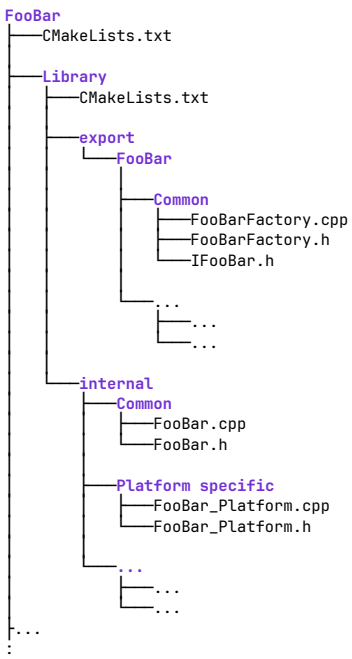


Abb. 1: Neue Ordnerstruktur [1]

Arbeitsbeschreibung und Methodik

Für die genannte C++-Codebasis gibt es derzeit zwei große Maßnahmen, die eine bessere modulare Struktur ermöglichen sollen.

Neue Ordnerstruktur: Um Abhängigkeiten sichtbar zu machen, sollen alle einzelnen Projekte in eine neue Struktur gebracht werden (siehe Abbildung 1). Alles, was von anderen Modulen verwendet werden kann(z.B. Factories), steht in „export“, die interne Implementierung in „internal“. Abhängigkeiten werden dadurch noch nicht reduziert, sondern müssen explizit angegeben werden. Früher gab es keine Unterscheidung zwischen „export“ und „internal“ und es konnte alles aus einem anderen Modul verwendet werden, indem man den Source-Folder in der jeweiligen CMake-Konfiguration mit eingebunden hat. Für die neue Modulstruktur müssen auch die CMake-Konfigurationen neu erstellt und angepasst werden.

Neue Projektstruktur: Abbildung 2 zeigt einen Vergleich der Struktur und der Abhängigkeiten zwischen den einzelnen Komponenten im aktuellen und im zukünftigen Zustand. Eine bestimmte Anwendung (violett dargestellt) besteht aus verschiedenen Feature-Modulen (blau). Diese bestehen wiederum aus den Core-Modulen (orange). Im aktuellen Zustand verwenden die Features zusätzlich die Globale Feature-Konfiguration (rote Pfeile unten). Hier werden die Module passend für verschiedene Verwendungen konfiguriert. Das kann z.B. ein unterschiedliches Verhalten je nach Anwendung sein. Da sich dies alles an einem Ort befindet, kennt jedes Feature auch die Konfigurationen aller anderen Features. Daraus ergeben sich gegenseitige Abhängigkeiten (rote Pfeile), die zu unvorhersehbaren Effekten führen können.
Die geplante Struktur wird schrittweise für die einzelnen Features eingeführt. Dabei sollen die Features selbst nicht mehr „entscheiden“ können, wie sie sich verhalten sollen. Stattdessen wird dies von außen durch die Feature-Orchestrierung gesteuert, die von der globalen Feature-Konfiguration die passende Konfiguration erhält. Die Features sind also voneinander unabhängig und kommt hier nicht mehr unvorhersehbaren Fehlen bei eigentlich unbeteiligten Features.

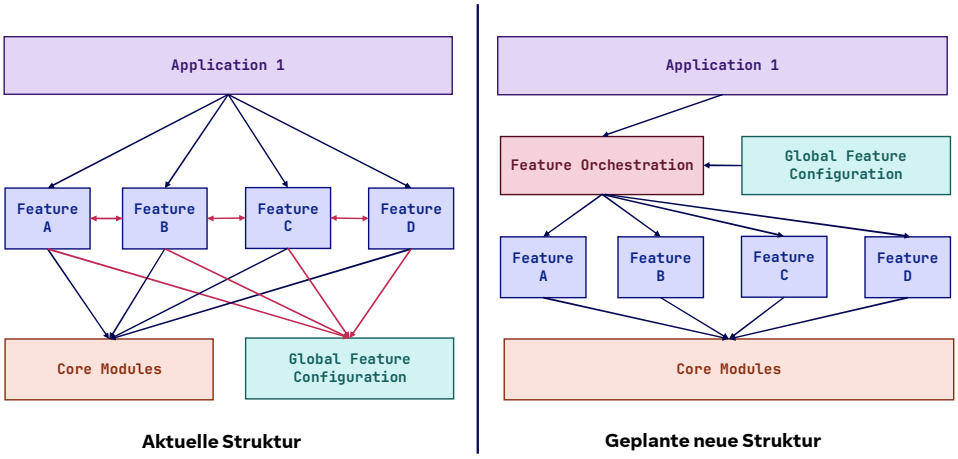


Abb. 2: Vergleich der Projektstrukturen [1]

Ausblick

Die Effektivität von Codeänderungen sollte mithilfe von Metriken gemessen werden. Für das Erstellen der Metriken über einen bestimmten Zeitraum ist die Verwendung eines Git-Checkout-Skripts geplant. Dieses betrachtet rückwirkend einen bestimmten Zeitraum, beispielsweise ein Jahr, indem es täglich Checkouts zu diesen Entwicklungsständen durchführt und diese miteinander vergleicht. Hierfür wird die Funktion „Sparse Checkout” verwendet, um gezielt nur einen bestimmten Teil des gesamten Codes (einzelne Dateien oder Ordner) betrachten zu können. Eine grundlegende Metrik ist das einfache Zählen der Feature-Definitionen, von denen es insgesamt Hunderte im Code gibt. Andere mögliche Metriken sind beispielsweise *Coupling Between Objects* (CBO) und *Lines of Code* (LOC) [3]. Diese Metriken sollen mit Diagrammen grafisch aufbereitet werden, um Trends oder auch Sprünge (z. B. durch große einzelne Refactorings) sichtbar zu machen. Das Ziel besteht darin, auf dieser Grundlage den Fortschritt der Modularisierung objektiv zu bewerten und zu zeigen, ob die Umstrukturierung messbare Verbesserungen erzielt hat.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Kitware Inc. CMake Documentation and Community. https://cmake.org/documentation, 2024.

[3] Alberto S. Nuñez-Varela, Héctor G. Pérez-Gonzalez, Francisco E. Martínez-Perez, and Carlos Soubervielle-Montalvo. Source code metrics: A systematic mapping study. *Journal of Systems and Software*, 128:164–197, 2017.

[4] Jesse Yli-Huumo, Andrey Maglyas, and Kari Smolander. How do software development teams manage technical debt?–An empirical study. *Journal of Systems and Software*, 120:195–218, 2016.

Entwicklung eines Frameworks für Clustering und Datenpartitionierung in Spring Boot-basierten Anwendungen

Nico Linder **Rainer Keller**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma AEB SE, Stuttgart

Einleitung

Moderne Softwarearchitekturen und Systeme zur Containerverwaltung wie Kubernetes ermöglichen es, Anwendungen schnell und flexibel zu skalieren, um auf wechselnde Lastverhältnisse zu reagieren und Redundanz zu gewährleisten. Für Softwareentwickler bringt die Umsetzung solcher verteilten Anwendungen jedoch auch Herausforderungen mit sich. Die Aufteilung einer Anwendung auf mehrere Instanzen macht es notwendig, diese zu koordinieren und Informationen wie Konfigurationsänderungen zuverlässig an alle Instanzen zu verteilen. Netzwerkprobleme oder ausfallende Instanzen sollen die Stabilität des Gesamtsystems nicht beeinflussen [4]. Zudem sind die Multithreading-Implementierungen von gängigen Sprachen wie Java oder Kotlin allein für die Verteilung von Rechenaufwand auf mehrere Anwendungsinstanzen oft unzureichend [2].

Typische Architekturmuster für verteilte Systeme wie Microservices ermöglichen zwar horizontale Skalierbarkeit und Redundanz, lösen aber nicht alle diese Probleme oder setzen die Integration zusätzlicher externer Komponenten für Aufgaben wie Lastverteilung und Koordinierung in jeder Anwendung voraus.

Für ein Softwareunternehmen, das bereits ein zentrales Framework nutzt, um verschiedene Softwareprodukte zu entwickeln, ist es zeitaufwändig und fehleranfällig, die Funktionalitäten für eine verteilte Anwendung in jedem Projekt von Grund auf neu zu implementieren. Stattdessen kann das bestehende Framework erweitert werden, um Entwicklern häufig benötigte Schnittstellen und Funktionen anzubieten, um ein solches Clustering umzusetzen.

Aktueller Stand und Motivation

Existierende Komplettlösungen für In-Memory-Datenspeicher und Computing-Plattformen bieten umfangreiche Funktionalitäten an, mit denen viele der zuvor beschriebenen Probleme gelöst werden können. Ein Beispiel für eine solche Plattform ist Hazelcast, diese wurde bereits in einem bestehenden Projekt der AEB SE eingesetzt. Dabei zeigte sich jedoch, dass die Verwendung einer teilweise proprietären Plattform als Komplettlösung erhebliche Risiken beinhalten kann. So besteht die Gefahr, dass benötigte Kernfunktionalitäten plötzlich nicht mehr oder nur noch unter proprietärer Lizensierung verfügbar sind. Zudem müsste die Plattform in jeder von der AEB SE entwickelten Anwendung, die Clustering benötigt, einzeln integriert werden, was einen erheblichen Mehraufwand darstellt. Ein weiterer Nachteil einiger verfügbarer Lösungen hinsichtlich der Parallelisierung von Aufgaben besteht darin, dass ihre Architektur darauf ausgelegt ist, die Ausführung von der eigentlichen Applikation zu entkoppeln und in spezielle „Ausführungs-Cluster“ auszulagern. Für das geplante Framework soll die Ausführung jedoch möglichst durch die Applikationsinstanzen selbst geschehen, externe Komponenten wie ein In-Memory-Datenspeicher sollen nur dazu dienen, die einzelnen Instanzen zu koordinieren, sowie in einem späteren Schritt gegebenenfalls für die Ausführung benötigte Daten innerhalb des Clusters zu replizieren.

Zielsetzung

Ziel dieser Arbeit ist es, ein Framework zu entwerfen, das häufig benötigte Funktionalitäten anbietet, um bestehende und neu entwickelte Softwarelösungen als Cluster aus mehreren Anwendungsinstanzen zu realisieren. Dafür soll das Framework die folgenden Funktionalitäten anbieten:

- Node Discovery bzw. Cluster Forming
- Leader Election und Möglichkeit, Aufgaben bestimmten Knoten zuzuweisen (Agreement-Protokolle)

- Mechanismus, um Nachrichten wie z.B. Zustands- oder Konfigurationsänderungen zuverlässig an alle Cluster-Mitglieder zu propagieren
- Schnittstelle, um Aufgaben zur Berechnung im Cluster zu verteilen
- Cluster-weites Rate Limiting bzw. Traffic Shaping, um durch hohe Parallelisierung z.B. aufgerufene Schnittstellen nicht zu überlasten

Der initiale Aufwand für Konfiguration und Nutzung der Clustering-Funktionalität soll für Entwickler so gering wie möglich gehalten werden. Für die Verteilung von Nachrichten, Daten und Aufgaben im Cluster muss sichergestellt sein, dass diese zuverlässig funktionieren, selbst wenn Knoten unerwartet wegfallen, neue Knoten hinzukommen oder kurzfristige Kommunikationsprobleme, beispielsweise durch Netzwerkfehler, auftreten. Die technische Umsetzung soll mit einer möglichst minimalen Schnittstelle zu externen Komponenten erfolgen, um die Abhängigkeit von speziellen Lösungen zu verringern und Drittanbieter-Software möglichst austauschbar zu machen.

Lösungsansatz

Die Open Source-Bibliothek Redisson bietet unter anderem verteilte Java-Datenstrukturen, Locking- und Sychronisationsmechanismen sowie einen Publish/Subscribe-Mechanismus, welche als Basis für die Implementierung der geforderten Funktionen dienen können. Soweit möglich sollen existierende Lösungen oder Algorithmen zur Umsetzung der Framework-Features so angepasst werden, dass die von Redisson angebotenen Funktionen und Datenstrukturen genutzt werden können. Redisson unterstützt als In-Memory-Datenspeicher unter anderem Redis, Valkey und Infinispan und dient als Abstraktionsschicht zwischen der Framework-Implementierung und der verwendeten Datenbank, wodurch die Abhängigkeit von einem bestimmten Anbieter verringert werden kann. Für die Koordinierung des Clusters können Agreement-Protokolle wie Two-Phase Commit (siehe Abbildung 1) verwendet werden, wobei beachtet werden muss, dass auch diese nicht immer komplett fehlerfrei funktionieren bzw. einen Kompromiss zwischen Zuverlässigkeit und Koordinierungsaufwand darstellen [3].
Für die Umsetzung sollen zunächst die Schnittstellen zum Entwickler definiert werden. Diese werden so konzipiert, dass die einzelnen Komponenten des Clustering-Frameworks einfach und unabhängig voneinander und von der zugrundeliegenden Implementierung verwendet werden können. Funktionen wie Leader Election, Cluster Forming oder Rate Limiting sollen direkt im Anwendungscode über Funktionsaufrufe oder Annotationen gesteuert werden können. Für die Verteilung von Aufgaben innerhalb des Clusters könnte ein Interface bereitgestellt werden, das der Entwickler implementieren muss, ähnlich der Multithreading-Implementierung von Programmiersprachen wie Java. Weiterhin muss evaluiert werden, ob die Verteilung über Warteschlangen, Locking-Mechanismen oder den von Redisson bereitgestellten Executor Service implementiert werden soll.

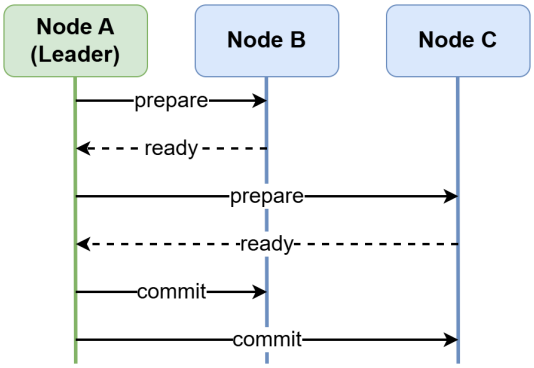


Abb. 1: Erfolgreicher Two-Phase Commit [1]

Um die Zuverlässigkeit des Frameworks sicherzustellen, müssen außerdem Tests definiert werden, welche die Eigenheiten verteilter Systeme berücksichtigen. Testmethoden hierfür beinhalten unter anderem das unerwartete Entfernen oder Hinzufügen von Knoten, simulierte Abstürze, Netzwerkfehler bzw. verlorene Nachrichten sowie Last- und Stresstests.

Ausblick

Das in dieser Arbeit entwickelte Framework soll eine einheitliche Grundlage für die Entwicklung verteilter Anwendungen innerhalb der AEB SE schaffen. Die darauf basierenden Anwendungen werden voraussichtlich über einen längeren Zeitraum im Einsatz sein, sodass Wartbarkeit und Erweiterbarkeit besonders im Fokus stehen. Eine mögliche Erweiterung des Funktionsumfangs wäre eine Schnittstelle, um von der Anwendung benötigte Daten für schnelleren Zugriff in verteilten Datenstrukturen innerhalb des Clusters zu speichern.
Um zu evaluieren, ob das Framework die zuvor formulierten Anforderungen erfüllt, soll ein existierender und von der AEB SE produktiv genutzter Service modifiziert werden, um zukünftig über mehrere Instanzen verteilt zu arbeiten.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Tauhida Parveen, Scott Tilley, et al. Towards a distributed execution framework for JUnit test cases. In *2009 IEEE International Conference on Software Maintenance*. IEEE, 2009.

[3] Michael Raynal and Mukesh Singhal. Mastering Agreement Problems in Distributed Systems. *IEEE Software*, 2001.

[4] Maarten van Steen and Andrew Tanenbaum. *Distributed Systems, 4th Edition*. Maarten van Steen, 2023.

Konzeption und Realisierung eines echtzeitfähigen Bus-Kopplungssystems auf Ethernet-Basis für moderne Ladesysteme

Friedrich Lohrmann **Walter Lindermeir**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma comemso electronics GmbH, Ostfildern

Einleitung

Im Zusammenhang der Klimaerwärmung und dem damit verbundenen Drang nach nachhaltigen und niedrigemittierenden Technologien hat die Elektromobilbranche in den letzten Jahren stark an Relevanz gewonnen. Trotz der momentanen Unsicherheit in der Automobilbranche ist der Sektor geprägt von ständiger Innovation und technologischem Fortschritt. Einen zentralen Bereich bildet hierbei die Kommunikation zwischen dem elektrischen Fahrzeug (auch *electric vehicle*, kurz: EV) und einer Ladesäule (auch *electric vehicle supply equipment*, kurz: EVSE). Hierfür wurden auf globaler Ebene viele Organisationen und Institute ins Leben gerufen, die Standards und Normen für Ladevorgänge in unterschiedlichen Systemen verwalten und weiterentwickeln, oder auch für neue Technologien neue Normen festlegen. Im Zuge dessen legen Automobilhersteller zunehmend einen Fokus auf frühzeitiges Testen bei der Entwicklung von neuen Elektromobilen. Dabei möchten sie sicherstellen, dass diese an möglichst allen verfügbaren Ladestationen geladen werden können, oder bei Abänderung einer Norm die Software in Fahrzeugen entsprechend anpassen, sodass diese wieder normkonform geladen werden können. Auf der anderen Seite wollen Ladesäulenhersteller sicherstellen, dass möglichst viele verschiedene Elektromobile an ihren Ladesäulen geladen werden können. Ein Resultat ist, dass eine große Nachfrage nach speziellen Testsystemen vorliegt, um sowohl EV als auch EVSE zu prüfen. Meist liegen diese Systeme als Schnittstelle zwischen den beiden Kommunikationspartnern vor, oder einer der beiden Partner wird bei Testvorgängen simuliert, um Kommunikation im Normal- und Störbetrieb zu analysieren [5]. Testgeräte basieren daher häufig auf leistungsstarken eingebetteten Systemen, die die Kommunikation überwachen, steuern oder simulieren. Sie bieten eine hohe Leistung und Echtzeitfähigkeit mit flexiblen Schnittstellen und einfach erweiterbarer Softwarelogik, die auch zukünftige Änderungen leicht integrieren können, um stets die neuesten Standards einzuhalten und zu testen.

Motivation und Ziel

Die eingesetzten Testsysteme sind meist so konzipiert, dass Geräte sowohl zu EV als auch zu EVSE eine Kommunikationsschnittstelle haben, falls beide Partner an Testfällen und Simulationen beteiligt sind. Im Zusammenhang dieser Arbeit wird ein eingebettetes System verwendet, welches am Microcontroller (kurz: µC) zwei Ethernet-Schnittstellen benötigt. Diese Schnittstellen verbinden den µC mit einem *physical layer transceiver* (kurz: PHY). Der PHY stellt eine Art Übersetzer dar, der Bits vonseiten des µC in elektrische Signale, die über ein Kabel verschickt werden können, umwandelt und umgekehrt [2]. Diese Anbindung erfolgt standardmäßig über ein *(reduced) media independent interface* (kurz: (R)MII). Der µC ist allerdings so konzipiert, dass er nur eine solche Anbindung unterstützt. Ziel der Arbeit ist es, zusätzlich eine zweite proprietäre Ethernet-Schnittstelle am µC zu implementieren, um einen effizienten und zuverlässigen Datenfluss im System zu gewährleisten.

ISO/OSI Modell

Das ISO/OSI Modell stellt eine allgemeine Referenz für Netzwerkprotokolle dar. Diese werden in Schichten dargestellt. Dabei liegen die unteren Schichten näher an der Hardware, während die oberen Schichten abstraktere Protokolle darstellen, die auf den unteren Schichten aufbauen [4]. Das Modell ist in Abbildung 1 abgebildet. Im Anwendungsfall der Arbeit wird das *transmission control protocol* (kurz: TCP) verwendet, um Daten zu verschicken. TCP-Segmente werden innerhalb von *internet protocol* (kurz: IP) Paketen übertragen, welche anschließend in einem Ethernet

Frame gekapselt werden, um eine Übertragung über das physikalische Netzwerkmedium zu ermöglichen.

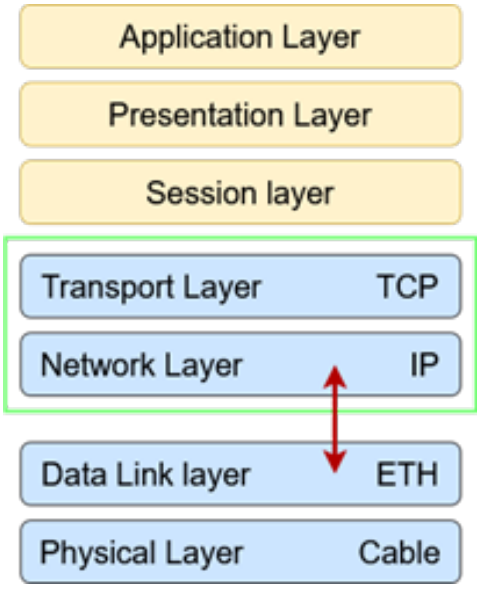


Abb. 1: ISO/OSI Schichtenmodell [1]

Umsetzung

Die erste der beiden benötigten Schnittstellen ist nach den Vorgaben des µC Herstellers implementiert. Für die Anbindung des verwendeten PHY über (R)MII werden bereits vorgefertigte Bibliotheken zur Verfügung gestellt. Auch für die Kontrolle und Konfiguration des PHYs zur Laufzeit werden Hardwarenahe Treiber für einige Modelle bereitgestellt. Da das Format der Basisregister und deren Modifikation genormt sind [3], konnte die Schnittstelle mit nur wenigen Änderungen an bereits bestehender Software implementiert werden. Eine zweite Ethernet-Schnittstelle am µC ist vom Hersteller nicht vorgegeben, daher musste die Anbindung des zweiten PHYs von Grund auf implementiert werden. Ein eigener hardwarenaher Treiber wurde implementiert, um Initialisierung, Steuerung und das Senden und Empfangen von Ethernet-Frames über den PHY zu ermöglichen. Für die Verarbeitung der Daten in höheren Schichten des ISO/OSI Modells kann eine Bibliothek des Herstellers verwendet werden, zu erkennen an dem Rahmen in Abbildung 1. In einem letzten Schritt wurden der hardwarenahe PHY-Treiber und Protokollstack noch verbunden, zu erkennen an dem Pfeil in Abbildung 1. Basierend auf der Softwarestruktur der ersten Schnittstelle wurden Funktionalitäten analog implementiert, sodass die Herstellerbibliothek den PHY-Treiber verwendet, um im Hintergrund der Anwendung automatisch über TCP eine Datenverbindung zu verwalten. Die Herstellerbibliothek stellt ein *application programming interface* (kurz: API) zur Verfügung, über deren Endpunkte Segmente leicht ausgelesen und verschickt werden können.

Evaluation

Für die Auswertung der Datenverbindung wurde in Windows Forms ein Evaluationsprogramm erstellt. Daten werden in einem Kreislauf von einem Laptop an den µC und wieder zurückgeschickt, dabei werden beide Schnittstellen durchlaufen. Das Programm zeichnet den Verbindungsstatus auf und testet, ob Datenpakete beide Schnittstellen erfolgreich durchlaufen oder verloren gehen. Zusätzlich werden einige Metriken wie Verlustrate oder die durchschnittliche Rundlaufzeit gemessen. Ein Ausschnitt der Oberfläche ist in Abbildung 2 zu sehen.

Abb. 2: Oberfläche des Analyseprogramms [1]

Aktueller Stand und Ausblick

Das grundlegende Ziel der Arbeit ist erreicht. Die Kommunikation über die Schnittstellen ist implementiert. Die Anwendung ist zusätzlich gegen Störeinflüsse wie Verbindungsabbrüche geschützt. Allerdings ist die Anwendung noch nicht in der Lage, größere Datenpakete von über 500 Byte ordnungsgemäß zu verarbeiten. Hier kommt es zu kritischen Fehlern, die nur durch einen Verbindungsabbruch und erneuten -Aufbau behebbar sind. Zusätzlich ist bei fehlgeschlagener Datenübertragung ein erneutes Senden erst nach minimal einer Sekunde möglich, da die Herstellerbibliothek keine kleineren Neuübertragungsintervalle zulässt. Das führt wie in Abbildung 2 rechts zu sehen zu teilweise sehr langen Rundlaufzeiten. Eine manuelle Anpassung der Herstellerbibliothek würde diese Zeiten verringern und nochmals die Reaktionsfähigkeit verbessern.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Jubilee Devi. Physical layer implementation of 1000BASE Ethernet, 2024.

[3] Institute of Electrical Engineers and Electronics. Telecommunications and Information Exchange Between Systems - Local and Metropolitan Area Networks - Specific Requirements Part 3: Carrier Sense Multiple Access with Collision Detection (CSMA/DC) Access Method and Physical Layer Specifications. https://ieeexplore.org/servlet/opac?punumber=4726157, 12 2005.

[4] Christian Facchi. *Methodik zur formalen Spezifikation des ISO/OSI Schichtenmodells*. Herbert Utz Verlag, 1995.

[5] Minho Shin et al. Building an interoperability test system for electric vehicle chargers based on ISO/IEC 15118 and IEC 61850 standards. *Applied Sciences*, 6:165, 2016.

Visualisierung von Explainable AI in der Kreditwürdigkeitsprüfung – Entwicklung und prototypische Umsetzung eines Analyse-Tools zur interaktiven Darstellung modellbasierter Vorhersagen

Patricija Loncaric **Giles-Arnaud Nzouankeu Nana**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

Der Einsatz von Künstlicher Intelligenz (KI) und insbesondere von maschinellen Lernverfahren (ML) hat in den letzten Jahren stark an Bedeutung gewonnen. In zahlreichen Bereichen werden automatisierte Entscheidungsprozesse durch datenbasierte Modelle unterstützt. [2] Besonders im Finanzsektor zeigt sich diese Entwicklung deutlich, da dort vermehrt Systeme zur automatisierten Bewertung der Kreditwürdigkeit eingesetzt werden. [1]
Diese Modelle analysieren historische Daten, um die Wahrscheinlichkeit eines Zahlungsausfalls vorherzusagen, und sind in der Lage, komplexe Zusammenhänge zu erkennen, die mit traditionellen statistischen Verfahren nur eingeschränkt erfassbar wären.

Problemstellung

Gleichzeitig birgt der Einsatz solcher Modelle erhebliche Herausforderungen im Hinblick auf deren Transparenz und Nachvollziehbarkeit und Fairness. Die Datenschutz-Grundverordnung (DSGVO) fordert etwa in Art. 22 die Begründbarkeit automatisierter Entscheidungen, sofern sie erhebliche Auswirkungen auf betroffene Personen haben. [5]
Entscheidungen wie die Ablehnung eines Kreditantrags sind weder für Antragstellende noch für Bankberater häufig nicht nachvollziehbar, insbesondere wenn sie auf Modellen wie Random Forests basieren (siehe Abbildung 1).

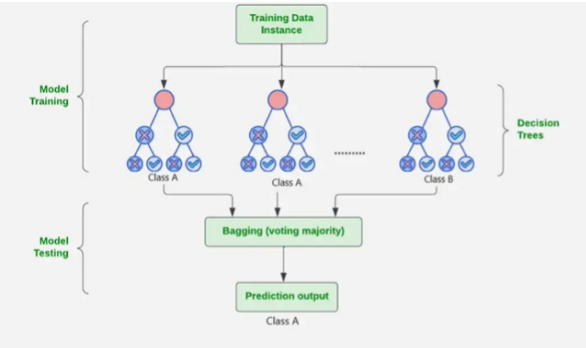


Abb. 1: Anwendung des Random-Forest-Algorithmus im Kreditwürdigkeitsprüfungsverfahren [6]

Um diesem Mangel an Erklärbarkeit entgegenzuwirken, befasst sich der Forschungsbereich Explainable Artificial Intelligence (XAI) mit der Entwicklung von Methoden, die maschinelle Vorhersagen verständlich und nachvollziehbar machen.

Zielsetzung

Ziel dieser Arbeit ist die Entwicklung eines interaktiven Prototyps, der die Kreditentscheidungen eines Machine-Learning-Modells mithilfe von Methoden der erklärbaren KI (XAI) verständlich erklärt. Dazu dient SHAP (SHapley Additive exPlanations), ein auf spieltheoretischen Konzepten basierendes Verfahren. Es quantifiziert den Beitrag einzelner Merkmale zur Vorhersage eines Modells und macht so nachvollziehbar, wie bestimmte Entscheidungen zustande kommen. Optional kann LIME (Local Interpretable Model-agnostic Explanations) verwendet werden. Dabei handelt es sich um einen modellunabhängigen Ansatz, der einzelne Entscheidungen durch lokal vereinfachte Modellapproximationen verständlich erklärt. [4]

Über eine benutzerfreundliche Oberfläche können die einzelnen Eingabeparameter wie Einkommen oder Laufzeit verändern und in Echtzeit nachvollziehen, wie sich diese spezifische Änderungen auf die Entscheidung auswirken. Dabei werden die Erklärungen durch verschiedene Visualisierungen ergänzt, um die Einflussfaktoren anschaulich darzustellen und die Transparenz für unterschiedliche Zielgruppen zu erhöhen.

Projektumsetzung

Die technische Umsetzung erfolgt auf Grundlage des öffentlich verfügbaren „German Credit Data Set“, der strukturierte Informationen über Kreditnehmer enthält (z. B. Alter, Beruf, Kredithöhe, Einkommen). [3]

Abb. 2: Struktur und Merkmale des German Credit Risk Datensatzes [3]

Nach der Datenvorbereitung werden verschiedene ML-Modelle trainiert und validiert. Für die finale Integration wird das Modell mit der XAI-Methode SHAP verbunden. Das Tool wird mit dem Python-Framework Streamlit realisiert. Nutzende können darüber interaktiv Eingabewerte verändern und die resultierenden Vorhersagen sowie deren Begründung unmittelbar nachvollziehen.

Ergebnisse und Funktionen des Prototyps

Der entwickelte Prototyp erlaubt die Eingabe hypothetischer Kundendaten, wie sie im Rahmen eines Kreditantrags erhoben werden. Auf Basis dieser Eingaben erfolgt eine Bewertung durch das ML-Modell. Die zugehörige SHAP-Analyse stellt dar, welche Merkmale (z. B. Einkommen, Beschäftigungsdauer) die Entscheidung in welchem Maße beeinflusst haben – positiv oder negativ.
Die Einbindung eines Radar-Diagramms ermöglicht es, kritische Abweichungen vom akzeptierten Standardprofil intuitiv zu erkennen. Darüber hinaus erlaubt das System die Simulation alternativer Eingabekonstellationen, sodass Nutzende nachvollziehen können, unter welchen Bedingungen eine abgelehnte Anfrage in eine Genehmigung überführt worden wäre.

Fazit und Ausblick

Die Ergebnisse der Arbeit sollen verdeutlichen, dass der gezielte Einsatz von Methoden des Explainable AI zur Verbesserung der Transparenz und Nachvollziehbarkeit automatisierter Kreditentscheidungen beitragen kann. Durch die Integration interaktiver Visualisierungselemente wird die Interpretation der Modellvorhersagen entsteht ein Mehrwert für Anwendergruppen ohne tiefgehende technische Kenntnisse.
Für zukünftige Forschungsarbeiten bietet sich eine systematische Usability-Evaluation des entwickelten Prototyps mit realen Nutzenden, wie beispielsweise Kreditnehmenden oder Bankberatenden, an. Dadurch könnten weiterführende Erkenntnisse zur praktischen Anwendbarkeit, Verständlichkeit und Akzeptanz erklärbarer Entscheidungsunterstützungssysteme im Finanzkontext gewonnen werden.

Literatur und Abbildungen

[1] Bundesanstalt für Finanzdienstleistungsaufsicht BaFin. Wenn ein Algorithmus über den Kredit entscheidet. https://www.bafin.de/SharedDocs/Veroeffentlichungen/DE/Fachartikel/2023/fa_bj_2305_Algorithmen_Kreditvergabe.html, 05 2023.

[2] Michael I. Jordan and Tom M. Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349:255, 2015.

[3] o.V. Kaggle. German Credit Risk. https://www.kaggle.com/datasets/uciml/german-credit, 2017.

[4] Zoumana Keita. Explainable AI – Understanding and Trusting Machine Learning Models. https://www.datacamp.com/tutorial/explainable-ai-understanding-and-trusting-machine-learning-models, 05 2023.

[5] Rolf Schwartmann, Andreas Jaspers, Gregor. Thüsing, Dieter Kugelmann, and Michael Atzer. *DS-GVO/BDSG: Datenschutz-Grundverordnung, Bundesdatenschutzgesetz*. C.F. Müller, 3 edition, 2024.

[6] B. P. Upendra, K. Sattar, and A. Elngar. A Smart Irrigation System Using the IoT and Advanced Machine Learning Model. *Journal of Smart Internet of Things*, page 20, 2024.

Classification and Segmentation of Anomalies in Industrial Image Analysis

Nils Luebben **Markus Enzweiler**

Department of Computer Science and Engineering, Esslingen University

Work carried out at Balluff GmbH, Neuhausen auf den Fildern

Introduction

Ensuring consistent product quality and preventing the downstream propagation of manufacturing flaws are critical objectives in modern industry. However, achieving this through automated visual inspection presents significant hurdles. Defects can be subtle, varied in manifestation, and continuously emerging in unexpected forms, signaling critical process deviations that may lead to substantial rework, waste, or product failure if undetected. Anomaly detection (AD) offers a computational approach to this problem, formally identifying patterns that deviate significantly from an established norm. Deep learning (DL) methodologies have become prominent in AD, primarily for their ability to learn complex representations directly from high-dimensional image data without manual feature engineering. While powerful, DL approaches often require substantial training data and computation and their inherent "black-box" nature can be detrimental to critical applications where interpretability is central. Transitioning DL-based AD into effective industrial solutions uncovers further practical difficulties. A core issue is data reality: while images of normal products are often abundant, compiling a comprehensive dataset of all potential defect types is frequently impractical and costly due to their rarity and diversity. This scarcity can lead to model overfitting and makes robust performance validation challenging. Moreover, publicly available pre-training datasets often exhibit significant domain mismatch with specific industrial applications, limiting transfer learning efficacy. Industrial environments are also dynamic; ongoing changes in products, materials, or processes can cause distribution shifts, such as evolving defect appearances (covariate shifts) or the emergence of entirely new defect categories (prior probability shifts), necessitating continuous system adaptation and retraining. To navigate these multifaceted challenges, this work explores three complementary AD paradigms for Balluff Network Interface (BNI) components: supervised classification with weak localization using visual explainability methods, fully supervised semantic segmentation, and semi-supervised, feature-based anomaly detection learning primarily from normal data. Rather than identifying a single optimal method, this exploration systematically assesses their inherent trade-offs concerning detection accuracy, annotation requirements, localization quality, computational demands, and adaptability to data-scarce, evolving industrial settings.

CNN Classification with Grad-CAM Visualization

One approach investigated was binary classification, where Convolutional Neural Networks (CNNs) were trained to distinguish between 'normal' and 'defective' BNI components. To understand the CNN's decision-making for identifying defects and to achieve weak localization without pixel-level labels, Gradient-weighted Class Activation Mapping (Grad-CAM) was utilized [1]. The fundamental principle of Grad-CAM is to highlight image areas that most significantly contribute to a speficic layer by weighting each of the layer's feature map by its contribution to the classification outcome, called the neuron importance weights. The selection of the convolutional layer for Grad-CAM, typically one of the final ones, is strategic. Deeper layers in a CNN capture higher-level semantic visual constructs while still retaining crucial spatial information often lost in fully-connected layers. Let c be the target class and y^c its pre-softmax score. k indexes the chosen layer's feature map activations A^k. The neuron importance weights, α_k^c, are determined by computing the gradients $\frac{y^c}{A^k}$ and global average pooling them across their spatial dimensions to yield a single importance weight for each feature map:

$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}} \tag{1}$$

The Grad-CAM localization map is then computed as:

$$L^c = ReLU\underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{linear combination}} \tag{2}$$

A Rectified Linear Unit (ReLU) is applied to this weighted sum, zeroing out negative values and ensuring that the final map only highlights features that positively contribute to the prediction of the target class. The resulting heatmap, when upscaled and overlaid on the input image, visually highlights the regions the CNN deemed most influential for its classification, as shown in Figure 1.

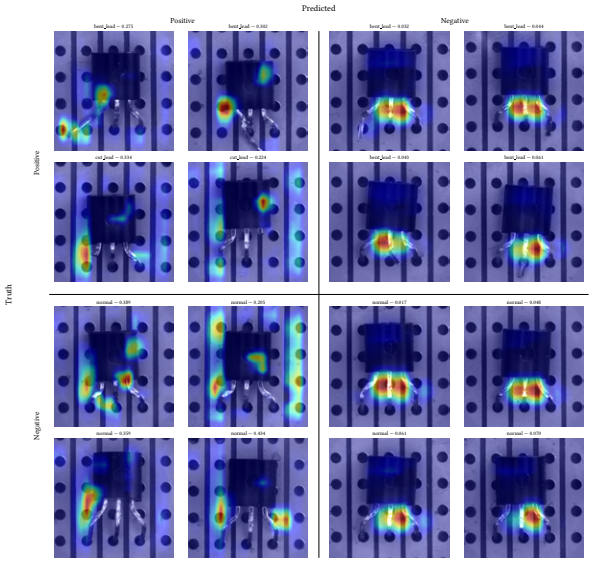


Fig. 1: Grad-CAM heatmap of transistor anomaly detection organized according to confusion matrix result combinations. [2]

Semantic Segmentation with U-Net

For pixel-precise defect localization, we implemented U-Net [3], selected for its strong performance with limited training data and robust localization capabilities. U-Net solves the fundamental trade-off between semantic context and spatial precision through its architecture. It consists of a contracting path that reduces spatial resolution to capture semantic features, and an expanding path that upsamples these features back to the original resolution (see Figure 2). The key innovation lies in the skip connections that link the contracting and expanding paths. These connections transfer high-resolution spatial details from early layers directly to the corresponding levels in the expanding path. This design ensures that the bottleneck focuses on encoding semantic context while the skip connections preserve essential spatial information. The expanding path then combines both inputs to determine what objects exist and where their boundaries lie. This architecture efficiently divides responsibilities between semantic understanding and spatial localization, enabling accurate pixel-level segmentation even with limited training data.

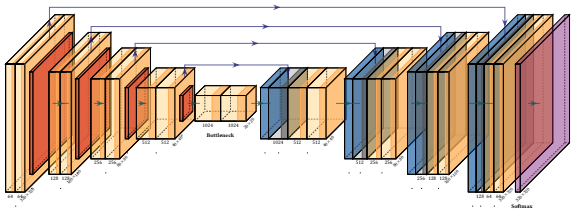


Fig. 2: U-Net architecture . Yellow blocks show convolutional layers, red blocks indicate max pooling operations, blue blocks represent transposed convolutions, blue arrows denote skip connections. [2]

Feature-Based Anomaly Detection with PatchCore

The third paradigm, PatchCore, is a feature-based anomaly detection method that relies on creating a maximally representative memory bank of nominal (normal) patch-features [4]. PatchCore utilizes a generic backbone CNN, pre-trained on a large dataset like ImageNet, as a fixed feature extractor; this backbone is not fine-tuned on the target industrial domain. Embeddings from several intermediate layers of the fixed, pre-trained backbone serve as patch features. Prioritizing earlier layers helps retain general visual information less specific to the backbone's original classes, improving robustness to distribution shifts in the target domain. These patch features, collected from all normal training images, initially form a comprehensive but large memory bank. Storing and querying this entire bank is resource-intensive. To address this, PatchCore employs coreset subsampling. This technique intelligently selects a significantly smaller subset of these initial patch features to create a compact yet highly representative final memory bank. The selection aims to ensure that the chosen subset effectively approximates the distribution of the full feature set, for instance, by iteratively picking features that are most distant from the already selected

coreset members, thereby maximizing coverage. This drastically reduces the memory footprint and inference time while largely preserving detection performance. During inference, a new test sample is processed through the same feature extraction pipeline to obtain its set of patch features. Each of these test patch features is then compared against all features stored in the coreset-reduced memory bank. The anomaly score for a test patch is typically determined by its distance (e.g., L2 distance) to its k-nearest neighbors (KNN) within the memory bank. An image-level anomaly score can then be derived, for instance, by taking the maximum anomaly score among all its patches. Furthermore, by spatially re-arranging these patch-level anomaly scores and upscaling them to the original image dimensions, an anomaly map can be generated, providing pixel-level localization of potential defects. Figure 3 showcases anomaly maps generated by PatchCore.

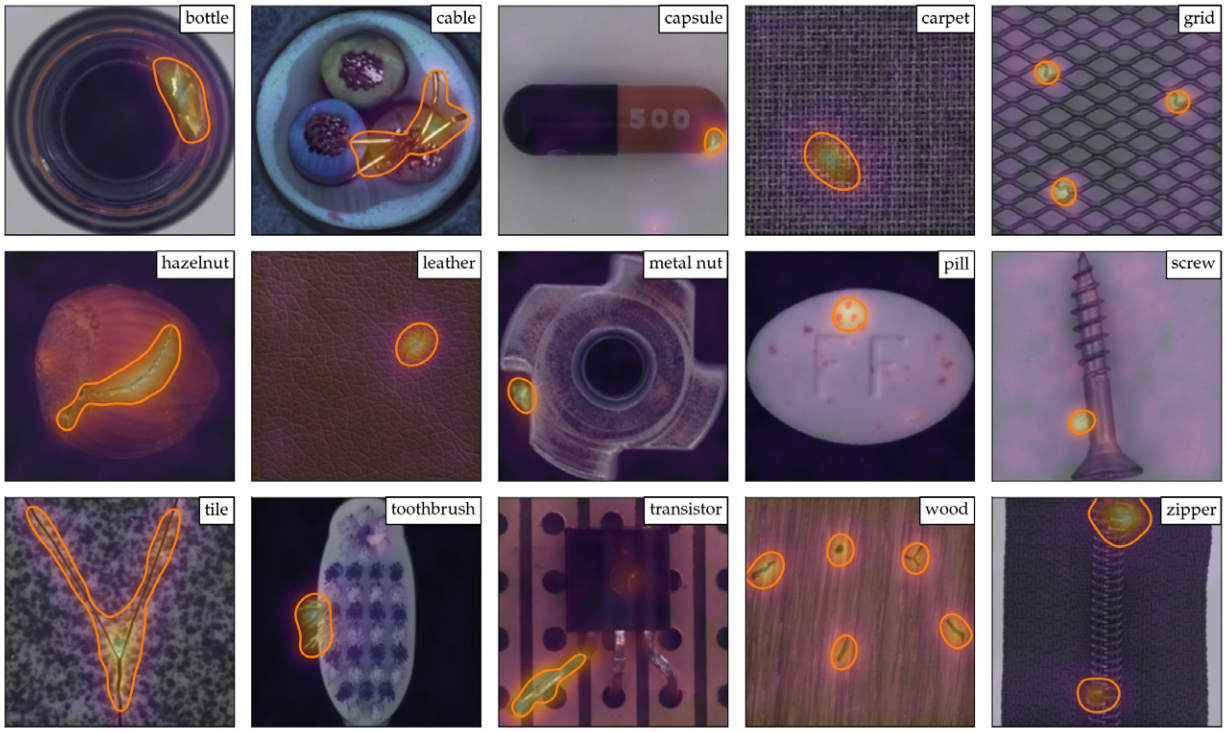


Fig. 3: PatchCore Segmentation Inference Samples from the MVTec benchmark datasets. The orange boundary denotes anomaly contours of actual segmentation maps for anomalies. [4]

Conclusion

This thesis evaluated three anomaly detection paradigms for industrial quality inspection, revealing distinct strengths and weaknesses. PatchCore provided the best image-level detection, U-Net excelled at pixel-level segmentation, and while classification models were fast, their Grad-CAM localization was often insufficient. The core insight is the necessity of navigating these trade-offs based on specific application needs. Looking ahead, a significant challenge and opportunity involve tailoring high-performing anomaly detection techniques for deployment on embedded systems with limited computational resources, a crucial step for practical industrial integration.

References and figures

[1] Selvaraju Ramprasaath et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017.

[2] Own representation.

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. https://arxiv.org/abs/1505.04597, 2015.

[4] Karsten Roth, Latha Pemula, Joaquin Zepeda, et al. Towards Total Recall in Industrial Anomaly Detection. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022.

Vergleich von KI-gestützten Code-Assistenten für Mainframe-Anwendungsmodernisierung – Ein wissenschaftliches technisches Nutzerreview des Watson Code Assistant for Z im Vergleich zu Konkurrenzlösungen

Leon Marquardt **Dirk Hesse**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma IBM, Böblingen

Einleitung

Die Entwicklung und Wartung komplexer Softwaresysteme – insbesondere im Mainframe-Umfeld – erfordert zunehmend intelligente Werkzeuge, um Effizienz, Codequalität und Wartbarkeit zu verbessern. In diesem Kontext gewinnen KI-gestützte Code-Assistenten an Bedeutung. Sie versprechen, die Produktivität von Entwicklerteams zu steigern, Fehlerquellen zu minimieren und den Entwicklungsprozess zu beschleunigen. Ein prominentes Beispiel ist der Watson Code Assistant for Z, der speziell für die Modernisierung von Mainframe-Anwendungen konzipiert wurde. Doch auch andere Anbieter bieten leistungsstarke Lösungen, die in unterschiedlichsten Entwicklungsszenarien zum Einsatz kommen. Die Herausforderung für Unternehmen und Entwickler besteht darin, den für ihre spezifischen Anforderungen geeignetsten Assistenten auszuwählen. Dabei stellt sich insbesondere die Frage: Welcher Code-Assistent unterstützt Entwickler am effektivsten dabei, ihre Ziele schnell und zuverlässig zu erreichen? Diese Arbeit widmet sich dieser Frage durch ein wissenschaftliches, nutzerzentriertes Review aktueller KI-gestützter Code-Assistenten. Auf Basis realitätsnaher User Stories und definierter Bewertungskriterien wird untersucht, welches Tool den größten Mehrwert in der Praxis bietet. Gleichzeitig entwickeln sich generative KI-Technologien rasant weiter. Laut Gartner werden bis 2025 rund 80% aller Softwareentwicklungsprozesse zumindest teilweise durch generative KI unterstützt [1]. Der IBM watsonx Code Assistant for Z adressiert dieses Potenzial gezielt für die Mainframe-Welt. Im Rahmen dieser Arbeit wird das Tool unter realistischen Bedingungen evaluiert, um seinen Beitrag zur Code-Verständlichkeit, Effizienzsteigerung und nachhaltigen Modernisierung zu bewerten. Die Arbeit erfolgt in Zusammenarbeit mit IBM und analysiert sowohl technische als auch methodische und wirtschaftliche Implikationen.

Technologischer Hintergrund

Der IBM watsonx Code Assistant for Z stellt ein KI-gestütztes Instrument zur schrittweisen Modernisierung von Mainframe-Anwendungen dar. Das zugrunde liegende Large Language Model (LLM) ist spezifisch für Mainframe Applikationen und Sprachen, wie COBOL, JCL, PL/I, und so weiter trainiert und unterstützt Entwickler in diversen Entwicklungsphasen, einschließlich der automatisierten Code-Analyse, Refaktorisierung und der Transformationsprozesse in moderne Programmiersprachen wie Java [2]. IBM beschreibt diese Methode im Rahmen einer schrittweisen Modernisierungsstrategie, die darauf abzielt, Risiken zu minimieren und gleichzeitig eine rasche Implementierung von Teilgewinnen zu ermöglichen [3]. Wie in Abbildung 1 dargestellt, gliedert sich der KI-gestützte Modernisierungsprozess bei IBM in die Phasen Understand, Refactor, Optimize, Transform und Validate.

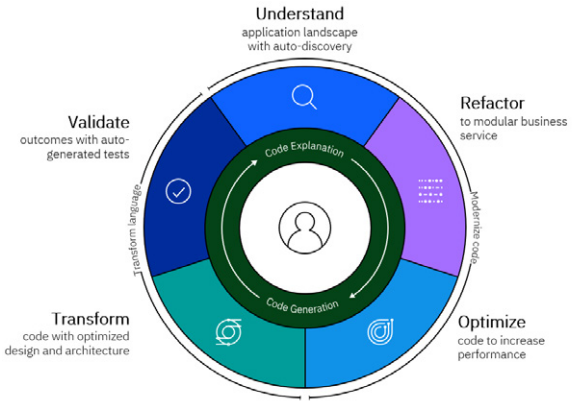


Abb. 1: Application lifecycle [2]

Nutzen und Anwendung des IBM watsonx Code Assistant for Z

Laut IBM konnte ein bedeutender europäischer Finanzdienstleister durch den Einsatz des IBM watsonx Code Assistant for Z den Aufwand für die Analyse und Umstrukturierung von COBOL-Code um bis zu 66% senken, was die Effizienz in den Bereichen Codeverständnis und Refaktorisierung erheblich steigert [2]. Westfield Insurance verzeichnete eine um 80% verkürzte Zeit zur Codeverständnis, was insbesondere beim Onboarding und bei Änderungsanforderungen große Vorteile bietet. Zusätzlich ermöglicht der Assistent eine automatisierte Generierung von Java-Code sowie von Unit-Tests, die semantisch äquivalent zum ursprünglichen COBOL-Quelltext sind. Dies erhöht die Nachvollziehbarkeit und reduziert das Risiko fehlerhafter Transformationen erheblich [4]. Wie in Abbildung 2 zu sehen ist, basieren diese Aussagen auf konkreten Fallbeispielen aus realen Kundenprojekten und sind Teil der Business-Value-Treiber, die IBM identifiziert hat.

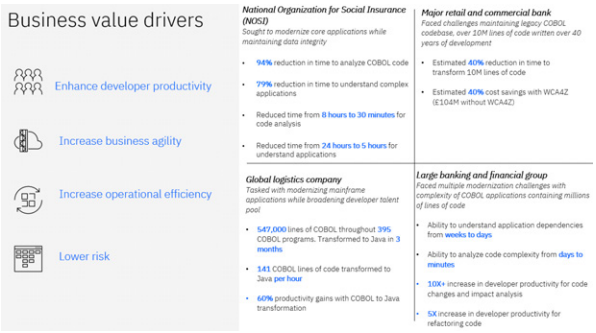


Abb. 2: Produktivitätssteigerungen bei Kundenprojekten [2]

Methodik der Bachelorarbeit

Ziel dieser Bachelorarbeit ist es, den IBM watsonx Code Assistant for Z mit führenden alternativen Code-Assistenten im deutschsprachigen Raum zu vergleichen und zu evaluieren, wie effektiv und nutzerfreundlich diese Tools typische Aufgaben im Kontext der Mainframe Application Modernization unterstützen. Dabei steht insbesondere der praktische Nutzen für Entwickler:innen im Fokus. Zur Zielerreichung werden folgende Teilziele verfolgt: ▪ Vergleich verschiedener KI-gestützter Code-Assistenten anhand praxisnaher Kriterien ▪ Analyse und Evaluation zentraler Funktionen im praktischen Einsatz ▪ Bewertung der Tool-Performance anhand realistischer User Stories ▪ Identifikation spezifischer Stärken und Schwächen des Watson Code Assistant for Z ▪ Ableitung konkreter Handlungsempfehlungen für zukünftige Verbesserungen und Weiterentwicklungen Die Untersuchung erfolgt entlang folgender Forschungsfragen: ▪ Welche Code-Assistenten ermöglichen es Entwickler:innen, ihre Ziele am schnellsten und effizientesten zu erreichen? ▪ In welchem Maß erfüllen die untersuchten Assistenten typische Aufgaben der Mainframe-Modernisierung? ▪ Welche Bewertungskriterien sind aus Nutzersicht besonders relevant für die Beurteilung von Code-Assistenten? ▪ Welche Alleinstellungsmerkmale und Limitierungen weist der Watson Code Assistant for Z im Vergleich zur Konkurrenz auf? ▪ Welche Rückschlüsse lassen sich aus den Erkenntnissen für die Weiterentwicklung KI-gestützter Code-Assistenten ziehen? Zur Beantwortung dieser Fragen werden verschiedene Tools – darunter der watsonx Code Assistant, GitHub Copilot sowie weitere Systeme – anhand realitätsnaher User Stories evaluiert. Es werden typische Aufgaben der Mainframe-Modernisierung exemplarisch umgesetzt, etwa das Erklären von COBOL-Code, die Generierung von Unit-Tests oder die Transformation in moderne Programmiersprachen. Die Bewertung orientiert sich an vordefinierter Kriterien wie Effizienz, Codequalität und Nutzerfreundlichkeit.

Erwartete Ergebnisse und Ausblick

Es wird prognostiziert, dass der watsonx Code Assistant signifikante Effizienzsteigerungen im Bereich der Code-Dokumentation und -Transformation im Vergleich zu herkömmlichen Werkzeugen bietet. Langfristig könnte die Integration einer Agentenarchitektur mit natürlicher Spracheingabe den Softwareentwicklungsprozess maßgeblich vereinfachen und beschleunigen. McKinsey prognostiziert, dass generative KI weltweit bis zu 4,4 Billionen USD an jährlichem Mehrwert erzeugen kann – ein erheblicher Teil davon im Bereich der Softwaremodernisierung [4]. Die in Abbildung 3 dargestellte Architektur zeigt exemplarisch, wie verschiedene KI-Agenten und ein zentrales Chat-Interface im Zusammenspiel Entwickler bei komplexen Aufgaben unterstützen können.

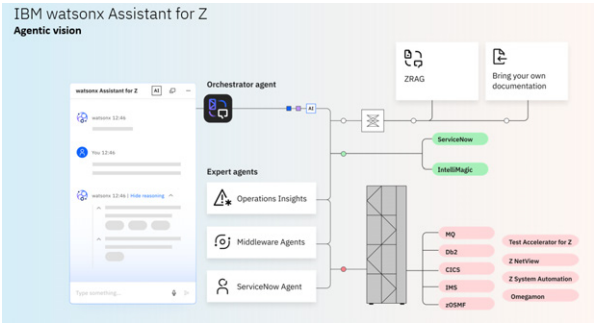


Abb. 3: Agentenarchitektur [2]

Literatur und Abbildungen

[1] Gartner Incorporated. Emerging Tech: Generative AI Code Assistants Are Becoming Essential to Developer Experience. https://www.gartner.com/en/documents/4348899, 2023.

[2] Pete McCaffrey. *watsonx Code Assistant for Z – Master Client Presentation*. IBM, 2024.

[3] IBM Research. watsonx Code Assistant overview. https://www.ibm.com/products/watsonx-code-assistant, 2023.

[4] Lareina Yee, Michael Chui, Roger Roberts, Eric Hazan, Alex Singla, Kate Smaje, Alex Sukharevsky, and Rodney Zemmel. The economic potential of generative AI: The next productivity frontier. https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier, 2023.

Entwicklung eines Wireless-Dongles zur drahtlosen Datenübertragung von Service- und Prozessdaten zwischen einem Brenner-Steuerungs-Systems und der dazugehörigen Anwendungssoftware

David Matussek **Michael Scharf**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Karl Dungs GmbH & Co. KG, Urbach

Einleitung

Das Unternehmen Karl Dungs GmbH & Co. KG entwickelt derzeit ein neues Brenner-Steuerungs-System für den Haushalt- und Industriebereich. Dieses System besteht aus mehreren Komponenten, die intern über einen CAN-Bus miteinander vernetzt sind. Der Zugang auf den internen CAN-Bus ist aus Gründen der funktionalen Sicherheit nur über eine USB-C-Schnittstelle an der Bedieneinheit möglich. Um das Brenner-Steuerungs-System in Betrieb zu nehmen, werden mit dem Bedien- und Inbetriebnahmewerkzeug, der sog. Toolbox, Parameter und Konfigurationsdaten über diese USB-C-Schnittstelle übertragen. Die Toolbox ist eine Anwendungssoftware, die auf handelsüblichen Endgeräten, wie bspw. einem Notebook oder Smartphone läuft. Für die Kommunikation wird ein proprietäres Protokoll namens ServiceLink verwendet, welches für die kabelgebundene Übertagung den USB-Gerätetyp Virtual COM Port nutzt. Das Werkzeug Toolbox realisiert hierbei einen ServiceLink-Client, während jede Komponente des Brenner-Steuerungs-Systems einen ServiceLink-Server implementiert.

Zielsetzung

Es soll für eine flexiblere Arbeitsumgebung gesorgt werden. Die Service- und Wartungsarbeiten am Brenner-Steuerungs-System sollen vereinfacht und mobilisiert werden. Dafür soll ein Wireless-Dongle entwickelt werden, der an der ServiceLink-Schnittstelle der Bedieneinheit angeschlossen wird und drahtlos mit der Toolbox kommuniziert. In dieser Bachelorarbeit wird untersucht, welche drahtlose Technologie und Bauteile sich hierfür eignen, sowie die prototypische Erstellung eines Wireless-Dongles inkl. der benötigten Software. Dabei sind sowohl technische Randbedingungen wie Stromversorgung, Größe und Datenrate, als auch Anforderungen wie Verschlüsselung, Reichweite, Kosten und die Einhaltung regulatorischer Vorgaben zu berücksichtigen. Abbildung 1 zeigt die grobe Funktionsweise des Projekts. Der Wireless-Dongle agiert an der Schnittstelle zu der Bedieneinheit als Virtual COM Port. Die Daten werden seriell gesendet und empfangen. Für die drahtlose Verbindung werden die Daten mit einem für die ausgewählte Technologie passenden Protokoll versendet oder empfangen.

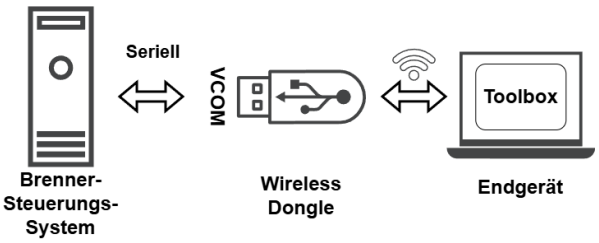


Abb. 1: Prinzipdarstellung [1]

Konzept

Abbildung 2 zeigt ein mögliches Konzept für den Wireless-Dongle, in der ein Funkmodul mit bereits integrierter Firmware eingesetzt wird. Dabei wird das Funkmodul von einem externen Microcontroller (Host-MCU) gesteuert. Zudem kommuniziert der Host-Controller per USB mit der Bedieneinheit des Brenner-Steuerungs-Systems. Das Konzept trägt sowohl zur Reduzierung des Aufwands als auch zur Vereinfachung der Produktzertifizierung bei. Zusätzlich werden weitere Bauteile für die Stromversorgung und die Störfestigkeit am USB-Port (EMC-Filter) benötigt. Der Entwurf bietet die Möglichkeit am Host-MCU weitere Peripherien, wie bspw. Speicher-ICs für die Ablage von Verbindungsdaten. Für das

Funkmodul kommen verschiedene Technologien infrage. In dieser Arbeit liegt der Fokus auf der Realisierung der drahtlosen Verbindung von WLAN oder Bluetooth, da diese Technologien weit verbreitet und in den meisten Geräten, auf denen das Bedienwerkzeug Toolbox ausgeführt werden kann, integriert sind.

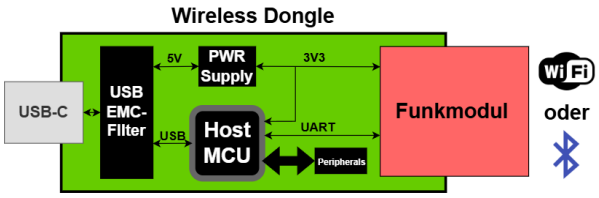


Abb. 2: Hardware-Entwurf Wireless Dongle [1]

WLAN

Beim Einsatz von WLAN wird in der Regel die Basic Service Set (BSS) Topologie verwendet. Diese beinhaltet einen Access Point (AP), der Geräten, auch Stations (STA) genannt, den Zugang zu dem Netzwerk gewährt. [4] Ein Funkmodul kann je nach Konfiguration als STA sich mit einem bestehenden Netzwerk verknüpfen, als AP ein Netzwerk aufbauen oder als STA und AP agieren. Aus Gründen der funktionalen Sicherheit wird in der Anwendung des Wireless-Dongles eine Einbindung in ein bestehendes Netzwerk ausgeschlossen. In Abbildung 3 wird ein mögliches WLAN-Netzwerk dargestellt. Darin agiert der Wireless-Dongle als STA und AP. Er baut ein Netzwerk auf, bei dem sich ein Endgerät mit der Toolbox als STA verbinden kann. Der Wireless-Dongle und das Endgerät können bspw. über TCP/IP miteinander kommunizieren. Hat das Endgerät eine Verbindung zum Dongle aufgebaut, kann es normalerweise nicht gleichzeitig in ein weiteres Netzwerk eingebunden sein.

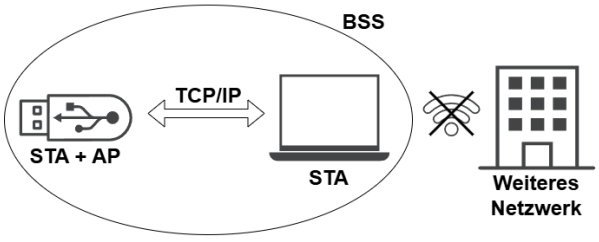


Abb. 3: WLAN-Netzwerk Konzept [1]

Bluetooth

Bei der Verwendung von Bluetooth wird ein Piconet-Netzwerk aufgebaut. Diese Topologie basiert auf einem Master–Slave Prinzip, das bis zu sieben Slaves aufnehmen kann. Dabei meldet sich ein Slave aktiv, sodass ein Master ihn erkennen und eine Verbindung herstellen kann. Zusätzlich gibt es eine Low Energy-Variante, die es ermöglicht, noch mehr Energie einzusparen. [3] In der Abbildung 4 wird ein Konzept für ein Bluetooth Netzwerk dargestellt. Dabei agiert der Wireless Dongle als Slave, sodass ein Endgerät mit der Toolbox als Master eine Verbindung herstellen kann. Dabei nutzen beide Geräte ein Serial Port Profile (SPP), um Daten seriell zu senden. Der Master kann zusätzlich sich mit weiteren Slaves verbinden und eine Anbindung zu einem weiteren Netzwerk über WLAN wäre weiterhin möglich. Aufgrund dieser zentralen Stärke wird das Bluetooth-Konzept bei der Entwicklung des Dongles berücksichtigt.

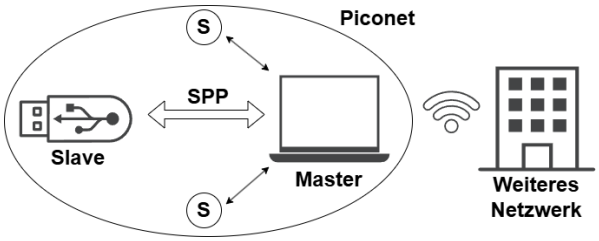


Abb. 4: Bluetooth Netzwerk Konzept [1]

Funkmodul

Als Funkmodul wurde der Stephano-I von Würth Elektronik eiSos GmbH & Co. KG ausgesucht. Der Stephano-I ist ein Kombi-Modul, das WiFi IEEE 802.11 b/g/n und Bluetooth Low Energy 5.0 unterstützt. Es ist mit einer von Würth entwickelten Firmware ausgestattet und verfügt über FOTA (Firmware Over-the-Air) Updates. Das Modul basiert auf dem Espressif ESP32 C3 Chip. Für die Steuerung der Verbindung und den Austausch von Daten kann ein Host MCU den Stephano-I mit AT-Commands über UART ansteuern. [5], [6] Die AT Commands sind kompatibel mit dem ESP-AT Projekt von Espressif. Dies fördert eine effektivere und schnellere Entwicklung. Um die WLAN und Bluetooth-Verbindung am Modul zu steuern, existiert eine Vielzahl an AT-Commands. [2] Die Vorteile eines Würth Produkts liegen in der Unterstützung der Zertifizierung und der Einbindung des Moduls in das Endprodukt, sowie der Verfügbarkeit einer ausführlichen Dokumentation.

Host - MCU

Die Host-MCU spricht das Funkmodul über eine UART-Verbindung mit AT-Befehle an und realisiert die USB-Device-Datenverbindung zu der Bedieneinheit

des Brenner-Steuerungs-Systems. Als Host-MCU wird ein STM32G0B1 ausgewählt. Dieser Microcontroller besitzt eine USB-Peripherie-Einheit und kann mithilfe des Konfigurationswerkezeugs STM32 Cube MX einfach initialisiert werden. Hierbei kann bspw. auf Bibliotheken für die Umsetzung eines VCOM Ports als USB-Device zurückgegriffen werden.

Ausblick

Auf Basis des Hardware-Konzepts muss für die Erstellung eines Prototypen als nächstes ein Schaltplan und ein PCB-Layout erstellt werden. Daraufhin wird die Software für den Host-MCU entwickelt. Dabei sorgt die Software für die Initialisierung und Steuerung des Funkmoduls über AT-Commands. Für die drahtlose Verbindung wird Bluetooth LE genutzt, da die Einbindung des Dongles in ein bestehendes Netzwerk ausgeschlossen wird. Daher wäre die Nutzung von WLAN kontraproduktiv, da sich das zu verbindende Endgerät dafür vom Arbeitsnetzwerk trennen müsste. Der Stephano-I bietet für Bluetooth LE ein Serial Port Profile ähnliches Profil, das generische Daten seriell senden und empfangen kann. Nach der Inbetriebnahme wird der Dongle getestet. Zunächst direkt am Brenner-Steuerungs-System auf Datenaustausch, Stabilität und Reichweite. Danach werden weitere Perfomance-Tests unter verschiedenen Umweltbedingungen, bspw. an einem echten Brenner, durchgeführt. Die Testergebnisse werden dabei protokolliert und anhand von Anforderungen bewertet, um Optimierung des Serienprodukts anhand des Prototypen vornehmen zu können.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Systems Espressif. What is ESP-AT. https://docs.espressif.com/projects/esp-at/en/release-v3.2.0.0/esp32c3/Get_Started/What_is_ESP-AT.html, 2016.

[3] Naresh Gupta. *Inside Bluetooth Low Energy*. Artech House, 2013.

[4] Tim Philipp Schäfers and Rico. Walde. *WLAN Hacking: Schwachstellen aufspüren, Angriffsmethoden kennen und das eigene Funknetz vor Hackern schützen*. Franzis Verlag, 2018.

[5] Elektronik Würth. *User Manual: Stephano-I*. Würth Elektronik eiSos GmbH & Co. KG, 2024.

[6] Elektronik Würth. Stephano-I & EV-Kits | Funkmodule | Würth Elektronik Produktkatalog. https://www.we-online.com/de/components/products/STEPHANO-I, 2025.

Barrierefreiheit im Web: Entwicklung eines Prototyps zur automatisierten Analyse und KI-gestützten Auswertung von Webseiten gemäß BITV 2.0

Niklas Meyer **Astrid Beck**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

Barrierefreiheit im Internet ist ein zentraler Bestandteil digitaler Teilhabe. In Deutschland ist die Barrierefreie-Informationstechnik-Verordnung (BITV) 2.0 die maßgebliche rechtliche Grundlage, die Anforderungen an barrierefreie Webangebote formuliert. Insbesondere öffentliche Stellen, aber auch private Akteure, stehen in der Verantwortung, ihre digitalen Angebote zugänglich zu gestalten. Dennoch bleibt unklar, in welchem Umfang große deutsche Webseiten diese Anforderungen umsetzen. Vor dem Hintergrund der zunehmenden Bedeutung des Internets im Alltag und der wachsenden Rolle von Künstlicher Intelligenz (KI) bietet die automatisierte Analyse von Barrierefreiheitskriterien eine Möglichkeit, strukturelle Defizite systematisch zu erfassen und vergleichend auszuwerten [3] [1].

Zielsetzung der Arbeit

Ziel dieser Arbeit ist es, mithilfe eines prototypischen Prüfsystems automatisiert zu erheben, ob und in welchem Maße große deutsche Internetseiten die Vorgaben der BITV 2.0 einhalten. Durch die Analyse mehrerer Dutzend Webseiten lassen sich Unterschiede zwischen Branchen erkennen sowie Aussagen darüber treffen, welche Aspekte barrierefreier Gestaltung gut umgesetzt werden und wo noch Verbesserungspotenzial besteht. Das System dient nicht nur der technischen Analyse, sondern soll perspektivisch als Werkzeug zur Sensibilisierung und Förderung digitaler Inklusion eingesetzt werden.

Bezug zur BITV 2.0 und geprüfte Kriterien

Die BITV 2.0 basiert auf den international anerkannten Richtlinien der Web Content Accessibility Guidelines (WCAG), stellt jedoch eine spezifisch deutsche Umsetzung dar. Der Prototyp deckt zentrale Prüfpunkte dieser Verordnung ab [3]. Die analysierten Kriterien umfassen:

- Formularelemente: Prüfung auf korrekt zugeordnete Labels. - Struktur: Analyse der Überschriftenhierarchie zur Sicherstellung einer logischen Gliederung. - Tabellen: Kontrolle auf semantisch richtige Tabellenstruktur.
- ARIA-Rollen: Bewertung, ob und wie zugängliche Rollen-Attribute verwendet werden. - HTML-Validität: Überprüfung mithilfe eines lokal gehosteten W3C Validators. - Sprache: Kontrolle auf vorhandene lang-Attribute sowie KI-gestützte Überprüfung der inhaltlichen Sprachkonsistenz. - Einfache Sprache: Einsatz von Sprachmodellen zur Einschätzung der Lesbarkeit und Verständlichkeit. - Bilder: Bewertung, ob alt-Attribute vorhanden sind und semantisch zur Bildaussage passen. - Kategorisierung: KI-gestützte Einteilung der Seiten in Inhaltsbereiche zur besseren Vergleichbarkeit.

Durch den Einsatz moderner KI-Dienste wie Google Gemini Vision können besonders inhaltsbezogene Merkmale — wie die semantische Qualität von Alternativtexten — objektiver bewertet werden.

Architektur des Prototyps

Der entwickelte Prototyp besteht aus einem Django-basierten Backend, das über RabbitMQ mit mehreren unabhängigen Testdiensten kommuniziert. Jeder Dienst übernimmt eine spezifische Aufgabe, z. B. das Parsen des HTMLs, das Extrahieren semantischer Strukturen, die Auswertung von Tabellen oder die Klassifikation von Textinhalten mittels KI.

Ein Architekturdiagramm veranschaulicht die Aufteilung der Komponenten sowie deren Kommunikationswege 1.

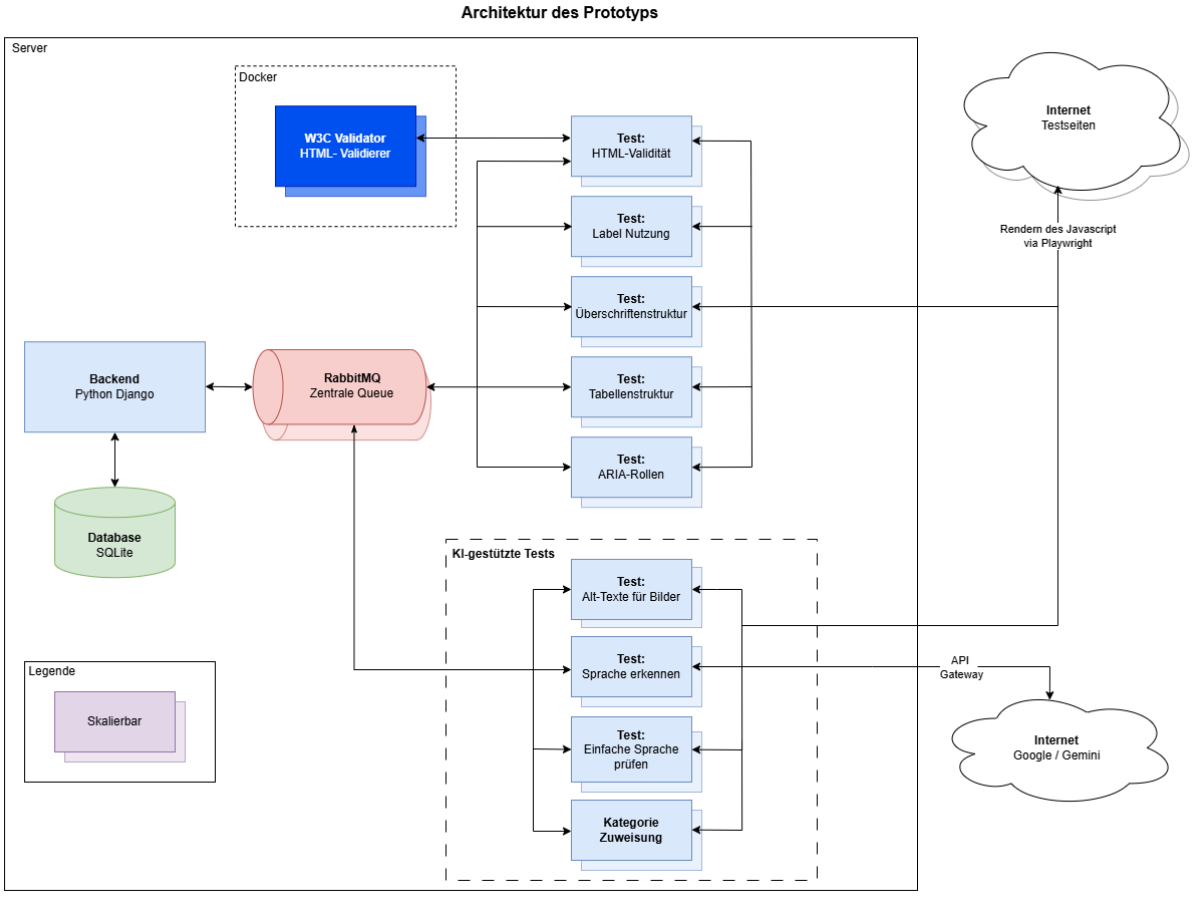


Abb. 1: Architektur des Prototyps [2]

Besonders hervorzuheben ist die Modularität, die zukünftige Erweiterungen, etwa durch zusätzliche Prüfkriterien oder alternative KI-Dienste, ermöglicht.

Beispiel eines Testablaufs

Ein beispielhafter Ablauf zur Analyse von Alternativtexten für Bilder beginnt mit dem Client, der über das Backend eine URL einreicht. Diese wird dem zuständigen Testdienst über RabbitMQ zugewiesen. Mithilfe von Playwright wird die Seite vollständig gerendert und analysiert. Die Bilddaten werden extrahiert, konvertiert und zusammen mit dem vorhandenen alt-Text an ein KI-Modell übergeben. Dieses bewertet die semantische Übereinstimmung und liefert eine Punktzahl zurück. Das Ergebnis wird schließlich gespeichert und an das Backend gemeldet.
Der genaue Ablauf ist im folgenden Sequenzdiagramm 2 dargestellt.

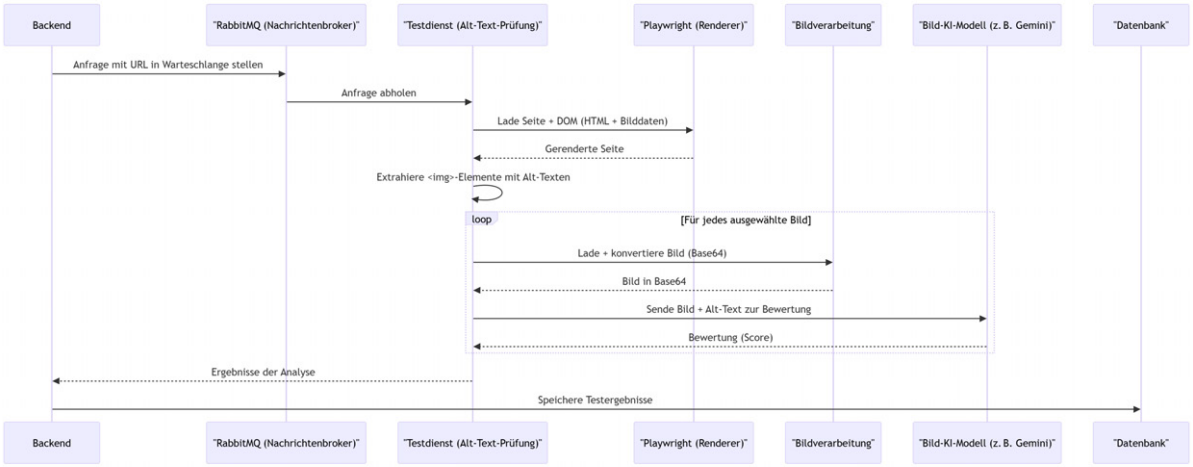


Abb. 2: Sequenzdiagramm eines Testablaufs [2]

Ausblick

Der aktuelle Prototyp stellt die Grundlage für ein erweiterbares System zur automatisierten Web-Barriereanalyse dar. Zukünftig wäre es denkbar, das System durch ein öffentliches Web-Frontend zu ergänzen, das es jedem Nutzer ermöglicht, eigene Webseiten auf Barrierefreiheit zu prüfen. Darüber hinaus könnten weitere Tests — z. B. zur Tastaturnavigation oder kontrastreichen Gestaltung — integriert werden.
Ein solches Werkzeug hätte nicht nur akademischen, sondern auch praktischen Mehrwert für Webentwickler, Unternehmen sowie öffentliche Stellen. Besonders im Zusammenspiel mit KI ergeben sich neue Chancen zur datengetriebenen Förderung von Barrierefreiheit [4].

Literatur und Abbildungen

[1] Marina Buzzi and Barbara Leporini. Is Wikipedia for the blind? https://dl.acm.org/doi/10.1145/1368044.1368049, 2008.

[2] Eigene Darstellung.

[3] Bundesministerium des Inneren und für Heimat. Barrierefreie-Informationstechnik-Verordnung – BITV 2.0. https://www.gesetze-im-internet.de/bitv_2_0/, 2023.

[4] Katrin Lang. *Auffindbarkeit, Wahrnehmbarkeit, Akzeptabilität*. Frank & Timme, 2021.

Methoden zur Kompensation von Geometriefehlern in industriellen CT-Anlagen

Alexander Moll **Thao Dang**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma HEITEC PTS GmbH, Kuchen

Einleitung

Mit dem zunehmenden Druck fehlerhafte Produkte so früh wie möglich im Produktionsprozess zu erkennen, bekommt die industrielle Computertomographie in den letzten Jahren eine immer größere Bedeutung in der Qualitätskontrolle. Im Gegensatz zur klassischen Röntgendurchstrahlungsprüfung wird in der Computertomographie ein Bauteil von mehreren Winkeln durchstrahlt und anschließend zu einem dreidimensionalen Volumen rekonstruiert. Eine Abweichung der vom Rekonsktionsalgorithmus angenommenen zu der tatsächlichen Aufnahmegeometrie führt zu Rekonstruktionsartefakten wie Doppelkanten und unscharfen Volumen, welche die nachfolgende Auswertung massiv erschweren oder sogar unmöglich machen. Industrielle Inline-CT Systeme stellen durch die Taktzeitvorgabe zusätzlich strenge Anforderungen an die Rechenzeit für Korrekturen, Rekonstruktion und Auswertung. Geeignete Geometriekorrekturverfahren müssen dadurch im Betrieb effizient und gleichzeitig langzeitstabil sein.

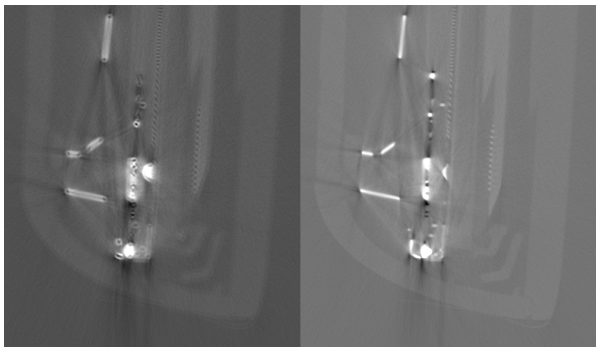


Abb. 1: Rekonstruktion ohne (links) und mit Geometriekorrektur (rechts) [1]

Ziel der Arbeit

Das Ziel dieser Arbeit ist die Untersuchung und Bewertung von verschiedenen Methoden um den Zeitaufwand für die mechanische Ausrichtung von industriellen Inline Kegelstrahl-CT-Systemen zu reduzieren und gleichzeitig die Bildqualität durch eine genaue Kenntnis der tatsächlichen Fehlstellung zu erhöhen. Der Fokus liegt dabei in einer möglichst universell einsetzbaren Lösung, die in unterschiedlichste Maschinengeometrien integrierbar ist und einen Vendor Lock-In vermeidet.

Fehlstellungsbestimmung

Geometrische Fehlstellungen können unterschiedliche Ursachen haben:

- Systematische Fehler durch Toleranzen (z.B. Fertigungs- und Montagetoleranzen)
- Systematische Fehler durch Positionierungsgenauigkeit (z.B. Robotersysteme)
- Stochastische Fehler (z.B. Rauschen)

Zur Bestimmung der geometrischen Fehlstellungen sind phantombasierte und phantomlose Verfahren üblich: Phantomlose Verfahren nutzen lediglich Projektionsbilder des eigentlichen Scans und führen Teilrekonstruktionen durch. In jedem Schritt wird dabei Bildqualität bewertet und die Fehlstellungsparamter iterativ optimiert [3]. Zusätzlich sind mit diesem Verfahren Erweiterungen möglich um stochastische Fehler während eines Scans zu erkennen und zu korrigieren [2] [4]. In phantombasierten Verfahren wird ein Messphantom (z.B. aus Stahlkugeln) in unterschiedlichen Positionen gescannt und die Positionen der Projektionen mit einem mathematischen Modell des CT-Systems mit allen möglichen Fehlstellungen verglichen. Das mathematische Modell wird durch einen Optimierer so lange angepasst, bis es mit den Messdaten übereinstimmt. Auf diesen Verfahren basierend wurden in den letzten Jahren viele verschiedene Verfahren und Algorithmen entwickelt, die für unterschiedlichste CT-Bauformen und Anwendungen optimiert sind [5]. Kalibrierte Phantome erlauben dabei eine bessere Abschätzung der Systemgeometrie in

Systemen mit vielen Freiheitsgraden (z.B. roboterbasierte C-Arm Systeme). Unkalibrierte Phantome sind dagegen günstiger und einfacher in der Handhabung, da keine hochpräzisen Fertigungsverfahren oder Vermessungen notwendig sind. Diese Verfahren berechnen die Phantomgeometrie selbst und stellen dabei größere Anforderungen an die Mechanik der Maschine. Phantombasierte Verfahren können bessere Ergebnisse als phantomlose Verfahren liefern, benötigen allerdings ein zu Maschine und Anwendung passendes Messphantom und einen Einmessvorgang. Phantomlose Verfahren sind dagegen flexibler, benötigen allerdings einen rechentechnisch deutlich höheren Aufwand zur Bestimmung der Fehlstellungen.

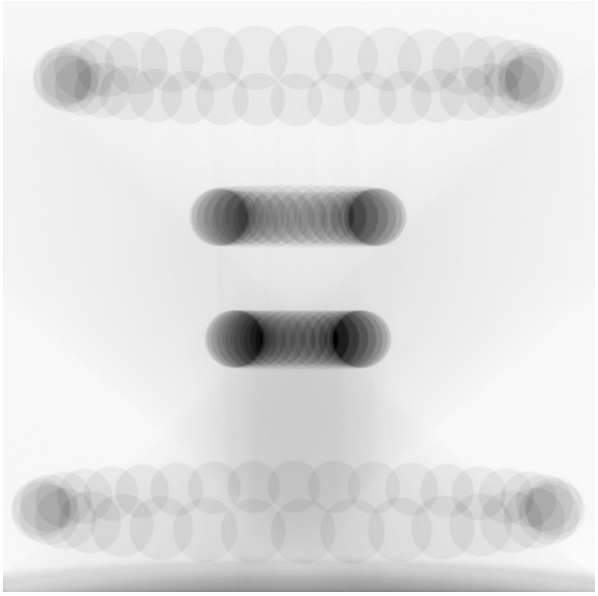


Abb. 2: Scantrajektorie eines möglichen Kalibrierphantoms [1]

Realisierung und Ausblick

Aufgrund der Vorteile durch geringe Kosten und Flexibilität im Aufbau von unkalibrierten Phantomen, sowie dem geringen Rechenaufwand wird dieser Ansatz als Grundlage für später folgende Optimierungen genauer untersucht. Dazu wurde im ersten Schritt ein CT-System mit allen möglichen Fehlstellungen mathematisch modelliert. Im weiteren Verlauf wurden unterschiedliche Kalibrierungsverfahren und Algorithmen basierend auf der Rotation eines unkalibrierten Phantoms aus Stahlkugeln implementiert. Unter Rücksichtnahme auf mögliche Maschinengeometrien und Anwendungen wurden anschließend Variationen der Messphantome bei unterschiedlichen Fehlstellungen simuliert und bewertet. Einzelne Variationen der Messphantome wurden zum Schluss real gebaut und die Geometrie einer realen Labor-CT vermessen. Dazu wurden die entsprechende Bildverarbeitung, Algorithmen und Simulationen in Python basierend auf NumPy, SciPy und OpenCV umgesetzt. Als weitere Schritte werden werden die Ergebnisse durch Vergleichsmessungen mit bestehenden Verfahren verglichen und bewertet. Im Anschluss muss die Reproduzierbarkeit der Ergebnisse unter verschiedenen Bedingungen an realen Maschinen getestet werden.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Siemens Healthineers. CERA - CT Geometry Calibration and Motion Compensation. https://www.oem-products.siemens-healthineers.com/software-components/ct-geometry-calibration, 2025.

[3] A. Kingston et al. Reliable automatic alignment of tomographic projection data by passive auto-focus. *Medical Physics*, 38:4934–4945, 2011.

[4] D. Rückert, L. Butzhammer, S. Wittl, G. Herl, T. Hausotte, and P. Kurth. Uncalibrated CT Reconstruction for One-Shot Scanning of Arbitrary Trajectories. In *13th Conference on Industrial Computed Tomography (iCT) 2023*, volume 29. e-Journal of Nondestructive Testing, 2024.

[5] Stefan Sawall, Michael Knaup, and Marc Kachelrieß. A robust geometry estimation method for spiral, sequential and circular cone-beam micro CT. *Medical Physics*, 39:5384–5392, 2012.

Erweiterung einer ML-basierten Unterstützungssteuerung zur Adaption eines mechanischen Hilfssystems zur Entlastung des menschlichen Körpers bei neuen Tätigkeiten

Jan Moser **Thao Dang**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma J. Schmalz GmbH, Glatten

Einleitung und Problemstellung

Aktive Exoskelette sind vielversprechende Technologien, um die Mobilität von Menschen in ergonomischen Arbeitsumfeldern zu unterstützen. Diese tragbaren Assistenzsysteme bieten eine physische Entlastung bei repetitiven oder belastenden Tätigkeiten. Durch die Verwendung von integrierter Sensorik und intelligente Steueralgorithmen können Exoskelette sich dynamisch an Bewegungen anpassen. Ein zentraler Bestandteil dieser adaptiven Steuerung ist der Einsatz von maschinellem Lernen (ML), das ermöglicht, Bewegungsmuster zu erkennen und individuelle Unterstützung zu leisten. Jedoch sind die ML-basierten Steuerungen in ihren Möglichkeiten begrenzt. Häufig sind diese Steuerungsalgorithmen auf vordefinierte Tätigkeiten trainiert und nur eingeschränkt übertragbar auf neue, nicht vorhergesehene Anwendungsfälle. Die Anpassung der Steuerung an neue Arbeitsumfelder ist derzeit mit hohem Aufwand verbunden. Sie erfordert eine umfangreiche Datenerhebung und das nachträgliche Training der Modelle in einem beaufsichtigten Umfeld. Dieser Prozess ist zeitaufwändig und hemmt die Flexibilität im praktischen Einsatz. Deshalb soll das Modell erweitert werden, um die Unterstützungssteuerung flexibel an spezielle Anwendungsfälle anpassen zu können. Durch die Integration von On-Device-Training können Exoskelette künftig selbständig lernen und sich an veränderte Bewegungsabläufe anpassen. Damit entsteht ein System, das nicht nur schneller einsatzfähig ist, sondern auch eine individuellere Unterstützung ermöglicht, ohne für jeden neuen Fall umfassende Trainingsdaten zu benötigen.

Konzept & Herangehensweise

Um die Anpassung der Unterstützungssteuerung an neue Anwendungsfälle zu ermöglichen, wird ein On-Device-Training auf der eingebetteten Recheneinheit des Systems realisiert. Die Motivation für eingebettete KI auf Mikrocontrollern liegt in der Anwendbarkeit, der Unabhängigkeit von der Netzinfrastruktur, der Sicherheit und dem Schutz der Privatsphäre sowie in den geringen Einsatzkosten [3]. Ziel dabei ist es, den Prozess direkt im Anwendungskontext durchzuführen – ohne externe Nachbearbeitung oder manuelles Nachtrainieren auf einem externen leistungsstarken Rechner. Das On-Device-Training ermöglicht dem Modell, sich an neue Daten anzupassen, indem ein vorab trainiertes Modell feinjustiert wird [2]. Die zentrale Idee basiert auf einem hybriden Modellansatz. Das Modell besteht aus einem vortrainierten Basis-Modell und einem adaptiven Modell, das on-device justiert werden kann, wie in Abbildung 1 dargestellt.

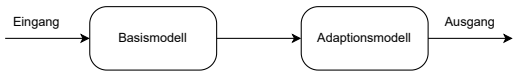


Abb. 1: Kombination von Basis- und Adaptionsmodell [1]

Das Basis-Modell ist dabei die Grundlage für die Verarbeitung der Sensorik und Vorhersage der Unterstützungskraft, basierend auf Sensoreingaben (z.B. IMUs, Kraftsensoren). Es ist auf typische Bewegungsmuster wie das Auf- und Abheben eines Objekts trainiert und bildet die Grundlage für die weitere Anpassungen des nachgeschalteten Modells. Falls das Modell des Exoskeletts im Anwendungsfall nicht korrekt reagiert, kann der Nutzer Korrektureingaben tätigen. Diese Interaktion erlaubt die Markierung von Fehlverhalten (z.B. „zu viel Unterstützung“ oder „nicht genug“), die als Korrekturhinweis genutzt werden. Das adaptive Modell nutzt die Rückmeldungen, um die Vorhersagen des Basismodells gezielt an den aktuellen Anwendungsfall anzupassen. Die Modifikation des Modells erfolgt ressourcenschonend, indem ausschließlich das angehängte Adaptionsmodell nachtrainiert wird. Ein besonderer Fokus liegt auf

der technischen Umsetzung unter eingeschränkten Hardware-Ressourcen. Für das Basismodell wurde eine Optimierung von FP64-Präzision (Floating Point 64-bit) auf FP32-Präzision (Floating Point 32-bit) durchgeführt, um Speicherverbrauch und Rechenlast auf dem Mikrocontroller zu reduzieren – bei möglichst geringem Genauigkeitsverlust. Das adaptive Modell wird mithilfe eines in der Programmiersprache C implementierten Frameworks umgesetzt. Das Framework stellt Funktionen für die Inferenz und Training eines neuronalen Netzwerks direkt auf Mikrocontroller bereit. Dabei können verschiedene Bausteine aus Schichten, Optimierungsalgorithmen und Verlustfunktionen ausgewählt und so ein Modell aufgebaut werden. Der Aufbau des ML-Modells unter Verwendung von dem Framework für das Adaptionsmodell wird in Abbildung 2 dargestellt.

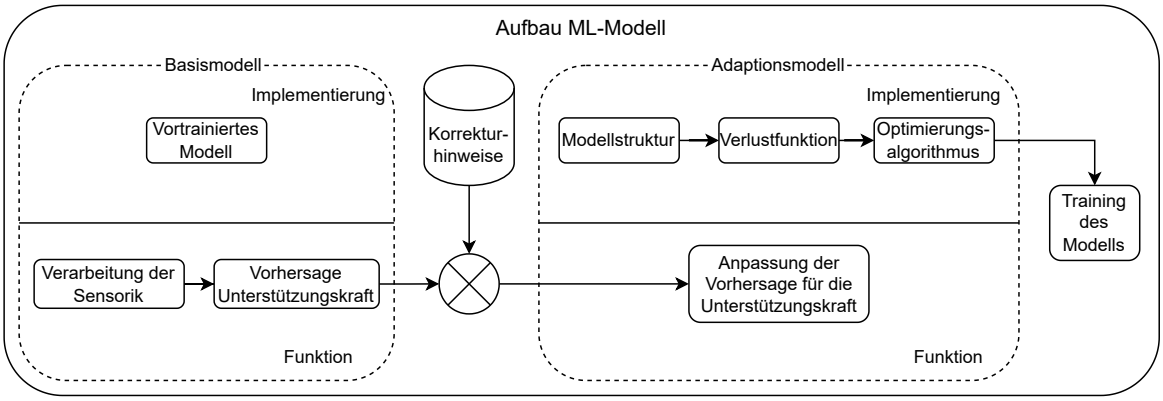


Abb. 2: Aufbau des ML-Modells [1]

Zusammenfassung und Ausblick

Die vorgestellte Herangehensweise bietet einen Ansatz zur Erweiterung einer ML-basierten Exoskelett-Steuerung durch On-Device-Training. Ziel ist es, die Bewegungssteuerung flexibel an neue Tätigkeiten anzupassen, ohne aufwendige Offline-Trainingsphasen. Die Kombination aus einem vortrainierten Basis-Modell und einem leichtgewichtigen, adaptiven Modell ermöglicht eine ressourcenschonende Feinjustierung direkt auf der Recheneinheit des Exoskeletts. Da zum aktuellen Zeitpunkt noch kein spezifisches Adaptionsmodell vorliegt, ist nur ein theoretischer Ausblick möglich. Dieser könnte wie folgt aussehen: Es werden verschiedene Modellansätze für die Adaption implementiert und trainiert. Dabei werden die verschiedenen Ansätze hinsichtlich der Modellgenauigkeit und der benötigten Rechenleistung auf dem Mikrocontroller ausgewertet. Mithilfe dieser Ergebnisse kann eine Prototyp-Implementierung mit einem realen Exoskelett durchgeführt werden. Außerdem können Nutzerstudien zur Evaluation der Anpassungsqualität und der Nutzerakzeptanz durchgeführt werden.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Ji Lin et al. On-Device Training Under 256KB Memory. https://arxiv.org/abs/2206.15472, 2024.

[3] Swapnil Sayan Saha et al. Machine Learning for Microcontroller-Class Hardware: A Review. In *IEEE Sensors Journal*. IEEE, 2022.

Integration von KI in die IT-Unternehmensarchitektur - Vorgehensweisen und Herausforderungen

Farhad Nessar **Manfred Schoch**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einführung

In jüngerer Vergangenheit hat sich die künstliche Intelligenz (KI) zu einem signifikanten Innovationsmotor der Unternehmens-IT entwickelt. Ihre Kompetenz, datengetrieben Entscheidungen zu treffen, Prozesse zu automatisieren oder neue Geschäftsmodelle zu erschließen, macht sie für Unternehmen aller Branchen zu einem hochrelevanten Faktor. Gleichzeitig führt ihre zunehmende Relevanz zu der Frage, wie sich KI-Technologien nachhaltig und strukturiert in bestehende IT-Unternehmensarchitekturen integrieren lassen – ohne dabei bestehende Systeme zu destabilisieren oder in Silos zu verfallen.

Forschungskontext und Zielsetzung

Etablierte Frameworks wie TOGAF stellen seit Jahren bewährte Orientierungsrahmen für IT-Unternehmensarchitekturen bereit. Hingegen mangelt es bislang an standardisierten Vorgehensmodellen für die Integration von KI. Die Herausforderung ist dabei doppelt gelagert: Einerseits sind auf technischer Ebene neue Anforderungen, wie etwa die Qualität der Daten, die Infrastruktur und die Governance, zu berücksichtigen. Andererseits sind auf organisatorische Seite Change-Management und strategisches Alignment erforderlich. Die vorliegende Bachelorarbeit adressiert diese Lücke, indem sie typische Vorgehensweisen, Muster und Herausforderungen bei der Integration von KI-Systemen systematisch untersucht. Das Ziel der vorliegenden Arbeit besteht in der Entwicklung eines praxisorientierten Verständnisses dafür, auf welche Art und Weise KI konkret in bestehende IT-Unternehmensarchitekturen integriert werden kann. Die vorliegende Arbeit stützt sich auf ein umfassendes theoretisches Fundament – bestehend aus den Themen KI, IT-Unternehmensarchitektur, Integration von KI in die IT-Unternehmensarchitektur - und analysiert Fallstudien aus der Praxis sowie qualitative Experteninterviews mit Fachleuten aus Unternehmen. Die Entwicklung eines Orientierungsrahmens in Form von Handlungsempfehlungen ist das Ziel des vorliegenden Projekts. Dieser Rahmen soll als Leitfaden für künftige KI-Integrationsprojekte dienen.

Künstliche Intelligenz

KI entwickelt Algorithmen, die Aufgaben übernehmen, die zuvor menschliche Intelligenz erforderten – etwa Lernen oder eigenständiges Handeln. Besonders durch ihre Fähigkeit, Muster in großen unstrukturierten Datenmengen zu erkennen, gilt sie als Schlüsseltechnologie für datengetriebene Unternehmensprozesse. Im Zentrum stehen neuronale Netze, deren Struktur dem menschlichen Gehirn nachempfunden ist. Sie bestehen aus einem Input Layer, mehreren Hidden Layers und einem Output Layer. Jedes Neuron verarbeitet Eingangssignale über gewichtete Verbindungen und Aktivierungsfunktionen, um komplexe Muster – von einfachen Kanten bis hin zu Gesichtern - zu erkennen.

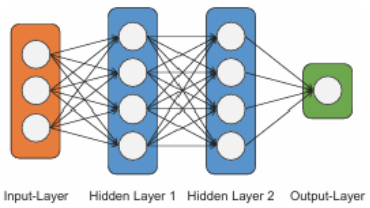


Abb. 1: Darstellung des Aufbaus eines neuronalen Netzes [6]

Diese Netze sind eng verknüpft mit dem Konzept des Maschinellen Lernens (Machine Learning), bei den Algorithmen aus Daten lernen, um eigenständig Muster zu erkennen und Entscheidungen zu treffen. Dabei unterscheidet man drei Lernformen: überwachtes Lernen mit vorgegebenen Beispielen, unüberwachtes Lernen zur Mustererkennung ohne Vorgaben und verstärkendes Lernen, bei dem das System durch Belohnung oder Bestrafung lernt [3]. Deep Learning

bildet eine Unterkategorie, bei der tiefe neuronale Netze mit vielen Hidden Layers eingesetzt werden. Sie sind in der Lage, relevante Merkmale direkt aus unstrukturierten Daten zu lernen – ohne manuelles Feature Engineering – und ermöglichen so hochgradig präzise Analysen und Vorhersagen [5].

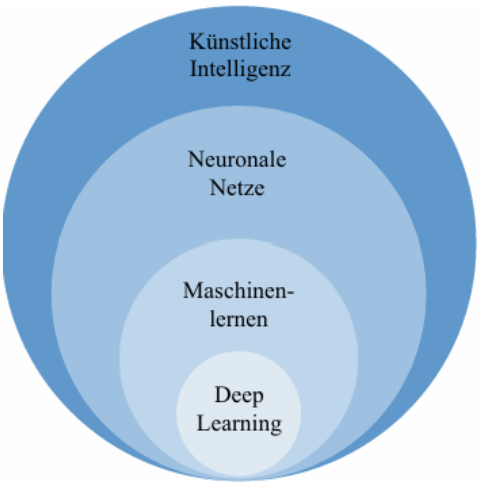


Abb. 2: Darstellung der Hierarchie der KI-Konzepte [4]

KI ist mittlerweile in nahezu allen Unternehmensbereichen im Einsatz. Im Kundenservice übernehmen Chatbots und virtuelle Agenten einfache Anfragen rund um die Uhr und entlasten so den Support. Machine-Learning-Modelle unterstützen bei der automatisierten Klassifikation von Tickets und der Generierung passender Antwortvorschläge. Im Marketing ermöglicht KI eine personalisierte Ansprache durch die Analyse von Nutzerdaten in Echtzeit, insbesondere über Empfehlungssysteme. In der Finanzabteilung kommt KI sowohl zur Automatisierung repetitiver Aufgaben als auch zur Entscheidungsunterstützung zum Einsatz – etwa bei Budgetierung oder Risikobewertung. Besonders im Bereich der Betrugserkennung bietet Deep Learning eine hohe Präzision bei der Identifikation komplexer Muster und Anomalien [7]. Auch in der Cybersicherheit analysiert KI-Netzwerkverkehr in Echtzeit, erkennt autonom Bedrohungen und leitet automatisch Schutzmaßnahmen ein – ein zentraler Beitrag zur IT-Sicherheit.

Enterprise Architecture

Die IT-Unternehmensarchitektur (Enterprise Architecture, EA) bildet den strategischen Ordnungsrahmen zur Ausrichtung der IT an den Unternehmenszielen. Sie umfasst alle relevanten Komponenten – von Geschäftsprozessen bis hin zur technologischen Infrastruktur – und gliedert sich typischerweise in vier Architekturschichten: Geschäfts-, Informations-, Anwendungs- und Technologische Architektur. Ziel ist es, durch strukturierte Planung, Dokumentation und Umsetzung eine durchgängige Transparenz und Effizienz in der IT-Landschaft zu erreichen [1].

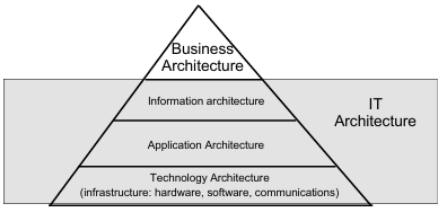


Abb. 3: Darstellung der Komponenten der Enterprise Architecture [1]

Das Enterprise Architecture Management (EAM) steuert die kontinuierliche Weiterentwicklung dieser Architektur. Es umfasst Aufgaben wie die Erfassung der Ist-Architektur, die Planung der Zielarchitektur sowie die Definition von Standards und die Sicherstellung der Governance. Zentrale Rollen sind u. a. der Enterprise Architect, Solution Architect und Demand Manager, ergänzt durch spezialisierte Architekten für jede Domäne. Auch Sicherheits- und Risikomanagement sind integraler Bestandteil. EAM steht jedoch vor Herausforderungen: manuelle Dokumentation, unklare Verantwortlichkeiten und fehlende Kommunikation führen häufig zu inkonsistenten Architekturen. Eine erfolgreiche EA-Strategie erfordert daher klare Prozesse, aktive Stakeholder-Kommunikation und eine starke Governance-Struktur [2].

Methodik

Die Arbeit stützt sich auf einem zweistufigen methodischen Ansatz, um zu untersuchen, wie KI konkret in die IT-Unternehmensarchitektur integriert wird. Zum einen wird eine qualitative Fallstudienanalyse durchgeführt, in der reale Anwendungsfälle systematisch ausgewertet werden. Dabei sollen zentral Vorgehensweisen, Muster und Herausforderungen identifiziert werden. Zum anderen werden ergänzend qualitative Experteninterviews eingesetzt, um die Erkenntnisse aus den Fallstudien durch praxisnahe Perspektiven der Unternehmenswelt zu erweitern. Beide methodische Zugänge dienen der Identifikation gemeinsamer Aspekte und Erfolgsfaktoren, die in Form von Best-Practice-Handlungsempfehlungen für Unternehmen mit KI-Integrationsvorhaben erläutert werden.

Erste Erkenntnisse & Ausblick

Die bisherige Analyse unterschiedlicher Fallstudien offenbart bereits gemeinsame Muster, die für die erfolgreiche Einbettung von KI in IT-Unternehmensarchitekturen von Relevanz sind. Dabei

zeigt sich als primäre Motivation die Effizienzsteigerung unternehmensinterner Prozesse durch KI-Einsatz – unabhängig von adressierten Branchen. Diese Zielsetzung lässt sich der Geschäftsarchitektur zuordnen, da sie unmittelbar auf die Gestaltung und Optimierung von Geschäftsprozessen abzielt. Des Weiteren lässt sich identifizieren, dass der Mensch in Verantwortung bleibt. Neben der gezielten Einschulung von Mitarbeitern und der Einbindung dieser in Entwicklungsprozessen bei einem Unternehmen, werden bei dem anderen Unternehmen Mitarbeiter verstärkt für die Qualitätskontrolle und Überprüfung der KI-generierten Ergebnisse eingesetzt. Diese Rollenverschiebung berührt insbesondere die Ebene der Geschäftsarchitektur, da sie bestehende Rollen- und Verantwortlichkeitsstrukturen verändert. Von besonderer Bedeutung ist die Integration von KI in die Informationssystemarchitektur. Die Fallstudien verdeutlichen, dass die Qualität der KI-Anwendungen maßgeblich von der Verfügbarkeit, Konsistenz und Zugänglichkeit relevanter Daten abhängt. Der Aufbau einer tragfähigen Informationsarchitektur bildet somit eine essenzielle Grundlage für den Erfolg der Integration. Zudem zeigen sich strategische Gemeinsamkeiten in der Herangehensweise: Unternehmen, die frühzeitig Machbarkeitsprüfungen durchführen und KI iterativ einführen, können potenzielle Risiken besser einschätzen und notwendige Anpassungen rechtzeitig vornehmen. Als häufigster Stolperstein erweisen sich dabei Datenverfügbarkeit und -qualität – ein Aspekt, der in den Experteninterviews weiter vertieft werden soll. Die identifizierten Muster werden im weiteren Verlauf der Arbeit durch die Experteninterviews validiert und zu praxisorientierten Handlungsempfehlungen verdichtet, das Unternehmen bei der Integration von KI in ihre IT-Architektur unterstützen soll.

Literatur und Abbildungen

[1] Bojidar Bojinov. THE ENTERPRISE ARCHITECTURE IN THE FIRM MANAGEMENT FRAMEWORK. *SSRN Electronic Journal*, 2016.

[2] Tim Bree and Erik Karger. Challenges in Enterprise Architecture Management: Overview and Future Research. *Journal of Governance and Regulation*, pages 355–367, 2022.

[3] Peter Buxmann and Holger Schmidt. *Künstliche Intelligenz - Mit Algorithmen zum wirtschaftlichen Erfolg*. Peter Buxmann & Holger Schmidt, 1 edition, 2019.

[4] Markus H. Dahm and Niklas Twesten. *Der Artificial Intelligence Act als neuer Maßstab für künstliche Intelligenz - Das Spannungsfeld zwischen Regulatorik und Unternehmen*. Springer Fachmedien Wiesbaden GmbH, 2023.

[5] John D. Kelleher. *Deep Learning*. The MIT Press, 2019.

[6] Ralf T. Kreutzer and Marie Sirrenberg. *Künstliche Intelligenz verstehen - Grundlagen - Use-Cases - unternehmenseigene KI-Journey*. Springer Fachmedien Wiesbaden GmbH, 2019.

[7] Teik Toe Teoh and Yu Jin Goh. *Artificial Intelligence in Business Management*. Kay Chen Tan & Dacheng Tao (Reihenherausgeber), 1 edition, 2023.

Entwicklung und Implementierung eines Frontend-Services zur Buchung und Überwachung von E-Scooter Sharing-Flotten

Huu Thinh Nguyen **Mirko Sonntag**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

E-Scooter sind mittlerweile ein fester Bestandteil moderner urbaner Mobilität und gelten als flexible und umweltfreundliche Lösung für kurze Strecken. Zahlreiche Anbieter ermöglichen den Zugang zu E-Scootern über benutzerfreundliche Apps, mit denen Nutzer verfügbare Fahrzeuge in ihrer Umgebung lokalisieren, reservieren und freischalten können. Informationen wie Batteriestatus, verbleibende Reichweite und Mietdauer ermöglichen eine gezielte und effiziente Nutzung. Durch die einfache Bedienung und die digitale Integration stellen E-Scooter eine attraktive Ergänzung zum öffentlichen Verkehr dar und fördern die Entwicklung nachhaltiger Mobilitätskonzepte. Diesen Ansatz verfolgt auch die Hochschule Esslingen und möchte ihr Mobilitätsangebot durch die Anschaffung von zunächst zwei neuen E-Scootern erweitern. Diese werden für Forschungszwecke am KEIM eingesetzt. Die beiden E-Scooter dienen für die Testphase zur Entwicklung einer neuen Software. In einem späteren Schritt ist geplant, die Anzahl der E-Scooter weiter zu erhöhen.

Zielsetzung

Ziel dieser Bachelorarbeit ist die Entwicklung einer benutzerfreundlichen App zur Buchung und Verwaltung von E-Scootern. Die App soll es ermöglichen, verfügbare Fahrzeuge schnell zu finden, deren Batteriestatus und Standort in Echtzeit abzurufen und eine Buchung einfach durchzuführen. Ein zentrales Anliegen dieser Arbeit ist zudem die Verbesserung der Mobilität innerhalb der Hochschule Esslingen. Die verschiedenen Standorte der Hochschule sind räumlich voneinander getrennt und oft nur mit erheblichem Zeitaufwand zu erreichen. Besonders für Studierende ohne eigenes Fahrzeug stellt dies eine tägliche Herausforderung dar. Durch den Einsatz von E-Scootern und der digitalen Unterstützung durch eine App soll eine flexible, zeitsparende und nachhaltige Lösung geschaffen werden, um die Wege auf dem Campus effizienter zu gestalten und den Hochschulalltag zu erleichtern.

Anforderung

Die entwickelte App soll eine Vielzahl von zentralen Funktionen bereitstellen, um eine komfortable und effiziente Nutzung der E-Scooter zu gewährleisten. Im Vordergrund steht dabei eine interaktive Kartenansicht, auf der alle verfügbaren E-Scooter in der Umgebung visualisiert werden. Zu jedem Fahrzeug sollen relevante Informationen wie der aktuelle Batteriezustand, der Standort sowie die Verfügbarkeit in Echtzeit abrufbar sein. Die Buchung und Freischaltung der E-Scooter erfolgt über einen QR-Code-Scan direkt an den Fahrzeugen und ermöglicht so einen schnellen Zugang. Darüber hinaus soll die App die aktuelle Position des Nutzers über das Smartphone erfassen, um eine präzise Navigation zu ermöglichen und definierte Parkzonen anzuzeigen. Innerhalb dieser Zonen kann der Nutzer seine Fahrt ordnungsgemäß beenden. Für den Betreiber werden zusätzliche Funktionen bereitgestellt. Dazu gehören das Anlegen und Anpassen von Parkzonen sowie die vollständige Übersicht über alle im Einsatz befindlichen Fahrzeuge. Darüber hinaus werden statistische Auswertungen über die Nutzung der E-Scooter zur Verfügung gestellt, um das Angebot kontinuierlich zu optimieren.

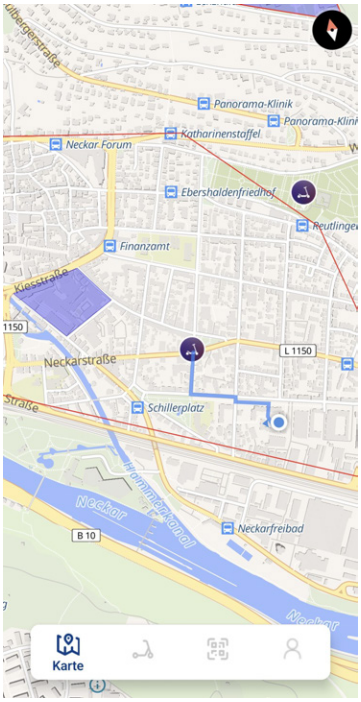


Abb. 1: Karten-Ansicht der App [1]

Nutzer zusätzliche Informationen wie Standort und Akkustand. Der Buchungsvorgang wird über das Scannen eines QR-Codes am jeweiligen Fahrzeug initiiert und abgeschlossen. In der Fahrtenverlauf-Ansicht kann der Nutzer seine bisherigen Fahrten einsehen, während in der Profilansicht persönliche Informationen sowie die Möglichkeit zur Abmeldung bereitgestellt werden.

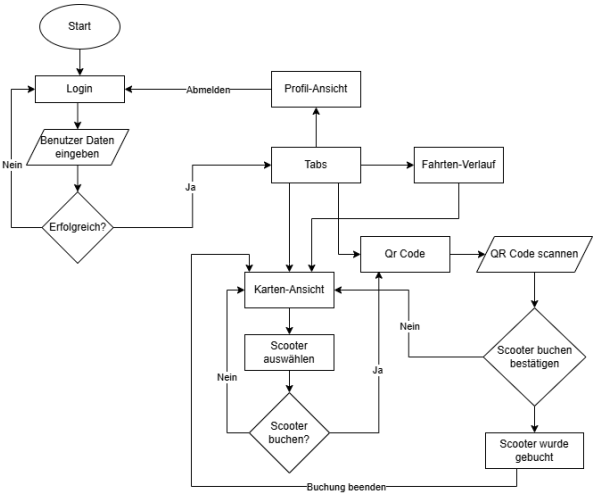


Abb. 2: Ablaufdiagramm der App [1]

Umsetzung

Die technische Umsetzung der E-Scooter App erfolgt mit dem Framework React Native in Kombination mit dem Framework Expo, um eine effiziente Cross-Plattform-Entwicklung für iOS und Android zu ermöglichen. Diese Technologie ermöglicht es, mit einer gemeinsamen Codebasis eine native Nutzererfahrung auf verschiedenen Endgeräten bereitzustellen. Expo erleichtert die Entwicklung erheblich, da es eine Vielzahl integrierter Tools und Services bietet - zum Beispiel für das schnelle und einfache Aufsetzen von Projekten, für einfaches Testen sowie für effizientes Debugging. Darüber hinaus bietet Expo eine große Auswahl an vorkonfigurierten APIs, die den Zugriff auf native Gerätefunktionen wie Kamera, Standort erheblich vereinfachen. [2]
Der Ablauf der App wird in Abbildung 2 dargestellt. Nach dem Start der Anwendung wird der Nutzer zunächst zum Login-Bereich weitergeleitet. Nach erfolgreicher Anmeldung gelangt er in die Hauptansicht der App, die über ein Tab-Menü den Zugriff auf vier zentrale Bereiche ermöglicht: die Kartenansicht, die Profilansicht, den Fahrtenverlauf sowie die QR-Code-Ansicht. In der Kartenansicht werden alle aktuell verfügbaren E-Scooter auf einer interaktiven Karte angezeigt. Beim Antippen eines Scooters erhält der

Ausblick

Für zukünftige Erweiterungen der Anwendung bieten sich verschiedene Funktionalitäten an, die sowohl die Nutzererfahrung verbessern als auch den Betrieb effizienter gestalten könnten. Ein attraktives Konzept besteht in der Integration von Gamification-Elementen. So könnte z.B. ein Punktesystem eingeführt werden, bei dem die Nutzer durch regelmäßige Nutzung, korrektes Parken in den Parkzonen oder das Melden von Problemen Punkte sammeln. Diese Punkte könnten gegen Freifahrten eingetauscht werden. Ein weiterer möglicher Entwicklungsschritt wäre die Einführung eines erweiterten Rollenkonzepts innerhalb der Anwendung. Neben den regulären Benutzern und Administratoren könnten dabei spezielle Rollen wie Ladepersonal oder Infrastrukturanbieter definiert werden. Diese Rollen hätten Zugriff auf zusätzliche Funktionen, beispielsweise zur Anzeige, Verwaltung und Wartung von E-Scooter Ladepunkten. Dadurch könnte die Versorgung und Instandhaltung der E-Scooter gezielter und effizienter organisiert werden. Ergänzend dazu wäre das Einbinden von E-Scooter Ladestationen in die Kartenansicht der App sinnvoll, sodass autorisiertes Personal die Standorte schnell identifizieren und anfahren kann.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Kuehl Hendrik. Vorteile von Expo und React Native in der App-Entwicklung. https://hybridheroes.de/blog/vorteile-expo-react-native-app-entwicklung/, 2024.

Einsatz von Künstlicher Intelligenz zur Verbesserung der Steuerung des Supply Chain eines Unternehmens

Khanh-Thien Nguyen **Giles-Arnaud Nzouankeu Nana**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

Künstliche Intelligenz (KI) ist heute aus unserem Alltag nicht mehr wegzudenken. Sei es als persönlicher Assistent auf dem Smartphone, zu Hause vor dem Fernseher, der passende Filme und Serien für den Abend empfiehlt, oder in Unternehmen, wo sie zum Beispiel Routinearbeiten automatisiert. In einer zunehmend digitalisierten Gesellschaft wird die Entwicklung und der Einsatz von KI immer relevanter – besonders für Unternehmen. Diese Arbeit beschäftigt sich mit dem Einsatz von KI im Supply-Chain-Management, insbesondere in den Bereichen Lieferanten- und Risikomanagement. Ziel ist es, die Potenziale von KI zu analysieren und praxisorientierte Ansätze aufzuzeigen.

Problemstellung

Die heutige Geschäftswelt ist sehr vernetzt und globalisiert. Unternehmen profitieren bei ihren Lieferketten von verbesserten Produktionsprozessen, dem Zugang zu Ressourcen aus dem Ausland und der Erschließung von Absatzmärkten. Doch müssen sie sich auch häufiger mit Gefahren auseinandersetzen. Die Lieferketten dieser Unternehmen sind anfälliger für eine Vielzahl von Störungen, die in letzter Zeit zugenommen haben. Naturereignisse wie extreme Unwetter oder geopolitische Instabilitäten und Konflikte verdeutlichen die Verwundbarkeit dieser Lieferketten. Neben diesen extremen Faktoren spielen auch unternehmensinterne Faktoren eine Rolle bei der Zuverlässigkeit der Lieferanten. Können Lieferanten ihre Abnehmer noch zuverlässig beliefern, obwohl eine Insolvenz droht? Im Rahmen dieser Arbeit besteht die Herausforderung darin, Industriekunden in die Lage zu versetzen, mithilfe eines KI-basiertes Systems Ausfallpotenziale früh zu erkennen, um vorzeitig Maßnahmen zu ergreifen. Ziel dieser Arbeit ist es, theoretische und praktische Ansätze zu entwickeln und eine Softwarelösung zu konzipieren, die mithilfe von Machine Learning frühzeitig Anomalien erkennt und dem Nutzer Entscheidungsinformationen bereitstellt.

Theoretische Grundlagen

Künstliche Intelligenz (KI): Die KI ist ein interdisziplinäres Forschungsfeld, das in den letzten Jahren große Fortschritte erzielt hat und dessen Technologie exponentiell wächst. Unternehmen wollen dieses Potenzial nutzen und investieren in entsprechende KI-Systeme. Es wird versucht, Systeme zu entwickeln, die eigenständig Probleme lösen oder mithilfe von Algorithmen aus großen Datensätzen lernen und daraus Muster erkennen können. Dazu gehören die Teilgebiete der KI, darunter die Künstlichen Neuronalen Netze (KNN), Machine Learning (ML) und Deep Learning (DL).

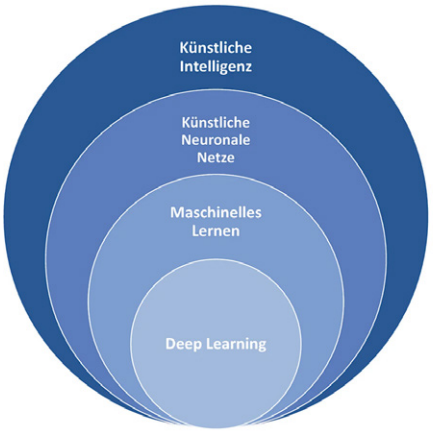


Abb. 1: Aufbau der Künstlichen Intelligenz [4]

Künstliche neuronale Netzwerke (KNN) sind durch das menschliche Gehirn und dessen Funktionsweisen inspiriert. Sie bestehen aus mehreren Schichten künstlicher Neuronen. KNNs werden bei der Bild- und Spracherkennung eingesetzt und dienen auch als Grundlage für das Deep Learning (DL), das wiederum zum Teilgebiet des Machine Learning gehört. Machine Learning (ML) bildet die Grundlage für KI-basierte Anwendungen. Mithilfe von Trainingsansätzen wie dem Supervised oder Reinforcement Learning sollen

komplexe Aufgaben wie Prognoseanalysen oder die Anomalienerkennung gelöst werden. [1]

Lieferantenmanagement: Innerhalb der Wertschöpfungskette nimmt die Beschaffung eine wichtige Rolle ein. Unternehmen wollen sicherstellen, dass die für die Produktion notwendigen Rohstoffe und Materialien effizient bereitgestellt werden. Mithilfe des Lieferantenmanagements soll dieser Bereich der Wertschöpfungskette sowohl in technischer als auch in wirtschaftlicher Hinsicht optimiert werden. Neben der Förderung der Abnehmer-Lieferanten-Beziehung dient das Lieferantenmanagement auch Prozessen zur Risikominimierung in der Lieferkette. [3]

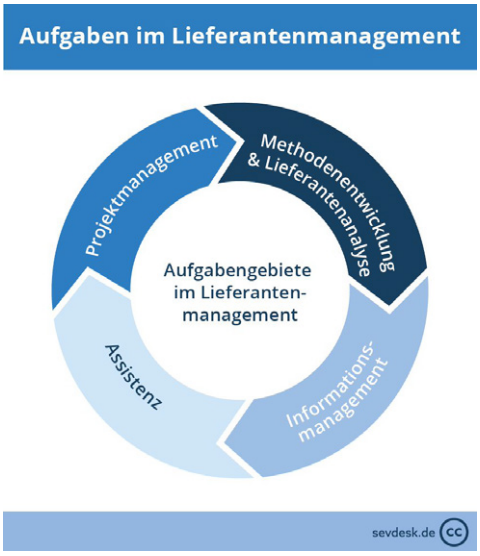


Abb. 2: Aufgabe Lieferantenmanagement [2]

Risikomanagement: Risikomanagement hat in seiner Gesamtheit die Aufgabe, Risiken zu identifizieren, zu beurteilen, zu steuern und zu überwachen. Das Risikomanagement in Unternehmen dient nicht nur der Schaffung von Möglichkeiten zur Risikominimierung, sondern soll auch Chancen aufzeigen, um diese nutzen zu können. Zu diesem Zweck werden Methoden und Tools wie z. B. das Scoring-System eingeführt.

Ausblick

Künstliche Intelligenz hat in letzter Zeit an Bedeutung gewonnen. Technische Fortschritte wie Echtzeiterkennung, Machine Learning und verbesserte KI-Modelle helfen Unternehmen, mithilfe von Predictive Analytics frühzeitig Entscheidungen in ihren Lieferantenbeziehungen zu treffen.

Literatur und Abbildungen

[1] Peter Buxmann and Holger Schmidt. *Künstliche Intelligenz - Mit Algorithmen zum wirtschaftlichen Erfolg*. Springer Gabler, 2019.

[2] sevDesk. GmbH. Aufgaben innerhalb des Lieferantenmanagement. https://sevdesk.de/blog/lieferantenmanagement/, 06 2024.

[3] Horst Hartmann. *Lieferantenmanagement - Gestaltungsfelder, Methoden, Instrumente mit Beispielen aus der Praxis*. Deutscher Betriebswirte-Verlag GmbH, 4 edition, 2019.

[4] Tom Hunger. *Mit künstlicher Intelligenz zu einer nachhaltigen Entwicklung - Eine qualitative Analyse auf Basis der Grounded Theory*. SpringerGabler, 2022.

Prototypische Entwicklung einer App mithilfe Compose Multiplatform

Dominik Oelke **Markus Enzweiler**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma pep.digital GmbH, Esslingen

Einleitung

Die plattformübergreifende App-Entwicklung stellt seit Jahren eine zentrale Herausforderung dar: Anwendungen sollen möglichst effizient auf Android, iOS, Desktop und perspektivisch auch im Web funktionieren – ohne separate Codebasen, redundante Logik des User Interfaces (UI) oder mehrfachen Wartungsaufwand. Compose Multiplatform (CMP) adressiert dieses Problem mit einem deklarativen, Kotlin-basierten Ansatz zur UI-Entwicklung für mehrere Plattformen auf Basis einer einheitlichen Architektur.

Technologischer Hintergrund

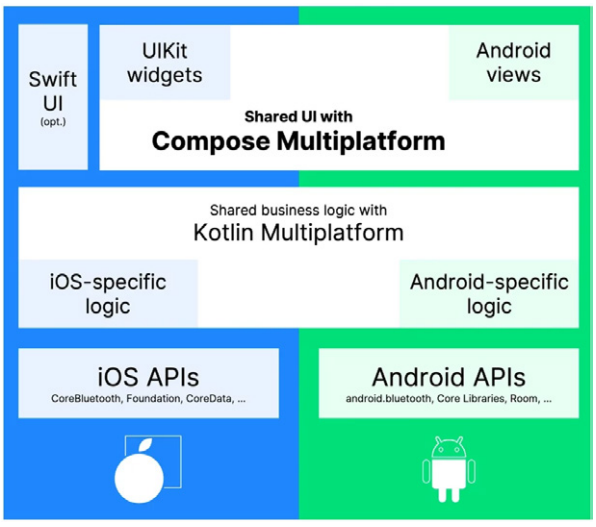


Abb. 1: Architektur von CMP am Beispiel von iOS und Android [1]

CMP kombiniert zwei Schlüsseltechnologien: Jetpack Compose, Googles UI-Toolkit für Android, sowie Kotlin Multiplatform (KMP), das die Wiederverwendung von Geschäftslogik über verschiedene Zielsysteme hinweg ermöglicht. Daraus ergibt sich ein Framework, das sowohl UI als auch Logik mit hohem Wiederverwendungsgrad in einer gemeinsamen Codebasis (siehe Abbildung 1) abbildet – bei gleichzeitiger Möglichkeit zur Integration plattformspezifischer Funktionen. [1]
Das deklarative Paradigma von Compose ermöglicht es, die Benutzeroberfläche aus dem aktuellen Zustand der Anwendung abzuleiten. Änderungen im Zustand führen automatisch zur Aktualisierung der UI, ohne manuelle Eingriffe. CMP erweitert dieses Modell um Desktop- und iOS-Unterstützung und bietet Interoperabilität mit bestehenden nativen Technologien wie SwiftUI, UIKit oder klassischen Android-Views. [2]

Architektur und Umsetzung

CMP-Projekte werden typischerweise modular aufgebaut: Ein zentrales Shared Module enthält die UI-Logik, während plattformspezifische Module für Android und iOS zusätzliche Implementierungen bereitstellen. Die Architektur folgt dem MVVM-Ansatz (Model-View-ViewModel) mit Unidirectional Data Flow (UDF). Zur Strukturierung kommen unter anderem Dependency-Injection (DI) (z. B. über Koin) sowie SQLite-basierte Datenhaltung zum Einsatz.
Die plattformübergreifende Nutzung erfolgt durch sogenannte expect/actual-Konstrukte: Gemeinsamer Code wird im Shared Module formuliert, während plattformspezifische Details – etwa Benachrichtigungen, Berechtigungsverwaltung oder System-APIs – in den jeweiligen Targets ergänzt werden. Für iOS erfolgt die Anbindung über Swift-Wrapper, wobei derzeit ein macOS-System für Builds vorausgesetzt wird.

Stärken und Schwächen im Überblick

Compose Multiplatform überzeugt insbesondere durch einen hohen Wiederverwendungsgrad der Codebasis.Ein Großteil der UI- und Logikkomponenten können plattformübergreifend implementiert werden, wodurch sich Entwicklungs- und Wartungsaufwand erheblich reduzieren lassen. Das deklarative UI-Modell ermöglicht

eine automatische Synchronisation der Benutzeroberfläche mit dem Anwendungszustand, was zu einer höheren Konsistenz und geringeren Fehleranfälligkeit führt. Durch die vollständige Integration in das Kotlin-Ökosystem – einschließlich Android Studio, Gradle und KMM – lässt sich CMP nahtlos in bestehende Entwicklungsprozesse einbinden. Auch das Plattformkonzept ist flexibel angelegt: Neben mobilen Zielplattformen wie Android und iOS werden ebenso Desktop-Systeme unterstützt.
Gleichzeitig sind einige Herausforderungen zu beachten. Die iOS-Unterstützung befindet sich derzeit noch im Beta-Stadium (Stand: Frühjahr 2025), was sich insbesondere bei komplexeren Anwendungen in Form von Einschränkungen bemerkbar machen kann. Auch die Tooling-Funktionalität ist noch nicht vollständig ausgereift: Lange Build-Zeiten und fehlende UI-Vorschauoptionen – insbesondere für iOS – erschweren die tägliche Entwicklungsarbeit. Hinzu kommt, dass für plattformspezifische Erweiterungen oftmals detaillierte Kenntnisse nativer Technologien erforderlich sind, etwa im Umgang mit Swift-Bindings oder systemnahen APIs.

Einsatzpotenzial und Perspektiven

CMP eignet sich insbesondere für Projekte mit begrenzten Ressourcen, die auf eine gemeinsame Codebasis für UI und Logik angewiesen sind – etwa in Startups oder kleinen Entwicklungsteams. Die Integration in bestehende Projekte ist durch die Interoperabilität mit nativen Komponenten erleichtert; auch inkrementelle Migrationen sind möglich.
Während Android- und Desktop-Ziele als stabil gelten, ist die iOS-Integration noch in der Beta-Phase. Die langfristige Eignung für produktive Multi-Plattform-Entwicklung hängt maßgeblich von der Weiterentwicklung durch JetBrains ab – insbesondere im Hinblick auf Web-Support, Tooling und iOS-Performance.

Zukunftsaussichten

Mit der kontinuierlichen Weiterentwicklung von CMP eröffnen sich neue Perspektiven für die plattformübergreifende Entwicklung moderner Benutzeroberflächen. Die geplante Stabilisierung der iOS-Integration sowie der Ausbau der Web-Unterstützung dürften CMP mittelfristig zu einem vollwertigen Cross-Plattform-Framework machen. Besonders relevant wird dabei die Verbesserung der Toolchain – etwa durch optimierte Build-Prozesse, konsistente UI-Previews und bessere Debugging-Optionen über Plattformgrenzen hinweg.
Parallel dazu wächst die Community rund um Kotlin Multiplatform, was langfristig zu einer verbesserten Dokumentation, einer breiteren Bibliotheksunterstützung und praxisnahen Best Practices führen dürfte. CMP könnte sich somit nicht nur als Alternative zu etablierten Frameworks wie Flutter oder React Native etablieren, sondern auch die Entwicklung nativer Oberflächen neu definieren – auf Basis eines konsistenten, deklarativen und vollständig in Kotlin integrierten Entwicklungsmodells.

Fazit

CMP stellt ein vielversprechendes Framework für die plattformübergreifende UI-Entwicklung dar. Es kombiniert die Vorteile deklarativer Benutzeroberflächen mit dem Potenzial einer wiederverwendbaren Kotlin-Codebasis. Besonders bei hoher UI-Komplexität, geringem Personaleinsatz oder begrenzter Projektzeit kann CMP helfen, Entwicklungsaufwände signifikant zu reduzieren. Für den breiten produktiven Einsatz wird es entscheidend sein, wie schnell die Plattformreife – insbesondere auf iOS und Web – weiter voranschreitet.

Ausblick auf weiterführende Untersuchungen

Für eine umfassende Bewertung von CMP wäre es sinnvoll, zukünftige Untersuchungen auf konkrete Vergleichsstudien mit anderen plattformübergreifenden Frameworks wie MAUI, Flutter oder React Native auszurichten. Dabei könnten Aspekte wie Entwicklungsaufwand, Performance, UI-Flexibilität, Barrierefreiheit sowie Benutzer- und Entwicklererfahrung systematisch gegenübergestellt werden. Auch der Einsatz in größeren, produktiven Multi-Plattform-Projekten bietet Potenzial für praxisnahe Erkenntnisse über Skalierbarkeit und Wartbarkeit.

Literatur und Abbildungen

[1] exyte droidcon. Jetpack Compose Multiplatform Android & iOS. https://www.droidcon.com/2023/07/17/jetpack-compose-multiplatform-android-ios/, 07 2023.

[2] JetBrains Writerside. Integration with the UIKit framework. https://www.jetbrains.com/help/kotlin-multiplatform-dev/compose-uikit-integration.html, 2024.

Entwicklung von Greifstrategien mit einer anthropomorphen pneumatischen Hand

Mahir Oezcan **Thao Dang**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Festo SE & Co. KG, Esslingen am Neckar

Motivation

Ein Griff ist eine statische Handhaltung, die es erlaubt, einen Gegenstand unabhängig von der Ausrichtung sicher zu halten [2]. In der Robotik ist die Fähigkeit Objekte zu greifen eine essenzielle Grundlage für sämtliche Manipulationsaufgaben. Viele Gegenstände im Alltag sind für menschliche Hände ausgelegt. Um menschenähnliche Geschicklichkeit beim Umgang mit diesen zu erreichen, empfiehlt sich der Einsatz anthropomorpher Roboterhände. Angesichts des zunehmenden Einsatzes von Roboter in dynamischen und unstrukturierten Umgebungen, wie etwas Krankenhäusern, ist der Einsatz menschenähnlicher Greifer unabdingbar. Trotz jahrzehntelanger Forschung bleibt das robuste Greifen jedoch ein offenes Problem. Die menschliche Hand besteht aus eine Vielzahl an Gelenken und Muskeln und besitzt tausende integrierte Sensoren. Nicht nur ist ein solch komplexes System schwer nachzuahmen, es erschwert zudem eine genaue Steuerung.

Zielsetzung

Das Ziel dieser Arbeit besteht darin, einer pneumatischen, anthropomorphen Hand das stabile und reproduzierbare Greifen verschiedenster Objekte beizubringen. Dazu sollen Deep-Reinforcement-Learning-Algorithmen wie beispielsweise Proximal Policy Optimization (PPO) verwendet werden. Im nächsten Schritt soll die in der Simulation erlernte Policy auf das reale System übertragen werden.

Related Work

Aufgrund des hochdimensionalen Aktionsraums und komplexer Systemdynamiken sind klassische analytische Methoden zur Steuerung oft nicht zielführend. In den letzten Jahren haben sich daher datengesteuerte Verfahren, allen voran Deep Reinforcement Learning, als vielversprechende Alternative erwiesen. Jüngste Arbeiten kombinieren RL mit vorab berechneten Greifposen [3] und großen Greifdatensätzen [6], um Griffe über zahlreiche Objekte hinweg zu lernen. In [7] wird ein Teacher-Student-Training genutzt, um zunächst objektspezifische Policies zu lernen, die via Distillation zu einer allgemeinen Policy zusammengeführt werden. Weitere Arbeiten wie [4] erweitern den Fokus vom reinen Greifen auf Pre-Grasp-Manipulation, d. h. vor dem eigentlichen Griff wird das Objekt aktiv in eine günstige Position gebracht.

Physisches System

Abb. 1: Der Anthropomorphic Soft Gripper [1]

Der Festo Anthropormphic Soft Gripper ist ein weicher Greifer der sich aus vier pneumatisch gesteuerten Fingern zusammensetzt, die um einen gemeinsamen Handteller angeordnet sind. Aller Finger sind von identischer Machart und bestehen aus je zwei Druckkammern $j \in 0, 1$, die sich bei Druckbeaufschlagung krümmen. Der Daumen (Finger 0) ist auf einem Schwenkantrieb befestigt und lässt sich auf einen gewünschten Winkel a_{S1} um die Handmitte rotieren. Diese Oppositionsbewegung ist essenziell für Präzisionsgriffe oder das Umgreifen größerer Objekte. Finger 1 und 2 sind jeweils auf Hebelzylindern (C1 und C2) befestigt und lassen sich so nach innen bzw. außen

schwenken. Finger 3 ist starr an der Basis monitiert. Insgesamt lassen sich über alle Aktuatoren hinweg elf Freiheitsgrade steuern. Die Hand besitzt keine Kontaktsensoren oder Kraftmessstreifen. Befestigt wurde der Endeffektor an einen Franka Research 3. Dabei handelt es sich um einen pneumatischen Roboterarm mit 7 Freiheitsgraden.

Modellierung

Die Entwicklung robotischer Kontrollstrategien, insbesondere im Bereich des Reinforcement Learnings, werden erst durch Simulationen praktikabel. Nicht nur lassen sich mehr Trainingsdurchgänge in kürzerer Zeit durchführen, sondern es werden auch potenzielle Schäden am Roboter verhindert. Um dies zu ermöglichen wird ein digitaler Zwilling des realen Roboters benötigt, der sämtliche physikalischen Eigenschaften des realen Systems abbildet. Die besondere Herausforderung bei einer Soft-Roboter-Hand liegt in ihrer hohen Elastizität und der damit verbundenen, praktisch unendlichen Anzahl an Freiheitsgraden, die eine direkte Steuerung erschwert. Analog zu früheren Arbeiten mit der Hand [5] wurde daher die Entscheidung getroffen, die Hand als Pseudo-Rigid-Body-Modell zu modellieren. Die Dynamik der Finger wird approximiert als eine offene kinematische Kette aus einer finiten Anzahl starrer Körper, die über Gelenke miteinander verbunden sind. So bleibt die Grundsätzliche Verformbarkeit erhalten, die Anzahl an Freiheitsgraden wird jedoch drastisch gesenkt. Abbildung 2 zeigt den Franka FR3 mit dem anthropomorphen Greifer in der MuJoCo-Simulation.

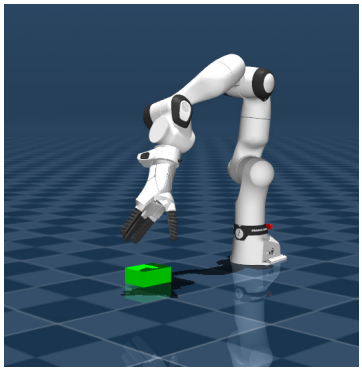


Abb. 2: MuJoCo-Simulation des Franka Emika Research 3 mit Festo Anthropomorphic Soft Gripper [1]

Ausblick

Im nächsten Schritt wird der Agent in der Simulationsumgebung darauf trainiert, ausgewählte Objekte zuverlässig zu greifen. Erste Tests mit einfachen Geometrien verliefen vielversprechend, jedoch müssen Trainingsprozess und Reward-Design noch finalisiert werden. Anschließend wird geprüft, wie gut sich die entwickelte Policy auf eine breitere Palette von Objekten übertragen lässt und ob Pre-Grasp-Manipulationen umsetzbar sind. Nach erfolgreichem Abschluss dieser Phase wird die erlernte Policy auf das reale System übertragen und dort validiert.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Thomas Feix, Javier Romero, Heinz-Bodo Schmiedmayer, Aaron M. Dollar, and Danica Kragic. The GRASP Taxonomy of Human Grasp Types. *IEEE Transactions on Human-Machine Systems*, 46:66–77, 2016.

[3] Shaowei Liu et al. ContactGen: Generative Contact Modeling for Grasp Generation. https://arxiv.org/abs/2310.03740, 10 2023.

[4] Dmytro Pavlichenko and Sven Behnke. Dexterous Pre-grasp Manipulation for Human-like Functional Categorical Grasping: Deep Reinforcement Learning and Grasp Representations. *IEEE Transactions on Automation Science and Engineering*, 2025.

[5] Gabriel Suske. Deep Reinforcement Learning für die Steuerung soft-robotischer In-Hand-Manipulationen, Masterarbeit, 2024.

[6] Ruicheng Wang et al. DexGraspNet: A Large-Scale Robotic Dexterous Grasp Dataset for General Objects Based on Simulation. https://arxiv.org/pdf/2210.02697, 03 2023.

[7] Wenbo Wang et al. UniGraspTransformer: Simplified Policy Distillation for Scalable Dexterous Robotic Grasping. https://arxiv.org/abs/2412.02699, 03 2025.

Einsatz von Künstlicher Intelligenz im Controlling – Status quo und Zukunftsperspektiven

Sila Pala **Catharina Kriegbaum-Kling**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

Künstliche Intelligenz (KI) gilt in vielen Unternehmensbereichen nicht mehr als Zukunftsvision, sondern wird zunehmend praktisch erprobt. Laut dem Statistischen Bundesamt nutzt etwa jedes achte Unternehmen in Deutschland KI-Technologien [1]. Verschiedene Anwendungen finden bereits heute in betrieblichen Prozessen statt, wobei das Interesse an KI-Lösungen weiter zunimmt. Auch im Controlling wird zunehmend diskutiert, inwieweit KI zur Analyse und Aufbereitung großer Datenmengen eingesetzt werden kann [4].
Studien belegen jedoch, dass bislang nur ein Teil der Unternehmen konkrete KI-Projekte umsetzt oder vorbereitet. Erste Einsatzbereiche betreffen vor allem vertriebs- und produktionsnahe Funktionen sowie Forschung, Entwicklung und Marketing [3]. Besonders kleinere und mittlere Unternehmen gehen vorsichtiger vor als große Betriebe, bei denen der Einsatz fortgeschrittener Technologien häufiger zu beobachten ist [1].
Mit dem wachsenden Interesse am KI-Einsatz steigen auch die Anforderungen an bestehende Strukturen. Die Fähigkeit, Daten systematisch zu erfassen, aufzubereiten und weiterzuverarbeiten, rückt stärker in den Fokus. Technische Voraussetzungen allein reichen für die Umsetzung nicht aus. Auch organisatorische Strukturen, rechtliche Rahmenbedingungen und interne Kompetenzen wirken dabei mit.
Neben technischen und strukturellen Aspekten spielen auch kulturelle und organisationale Bedingungen eine Rolle. In einigen Unternehmen bestehen Vorbehalte gegenüber der Technologie, während andere offener damit umgehen. Unterschiede zeigen sich auch in der Einschätzung: Führungskräfte bewerten die Potenziale tendenziell anders als Fachabteilungen, die näher an der praktischen Umsetzung stehen. Trotz dieser Unterschiede gilt KI in vielen Unternehmen als relevantes Zukunftsthema, vor allem mit Blick auf Prozessoptimierung, Informationsgewinn und die Möglichkeit, große Datenmengen effizienter auszuwerten [3].

Problemstellung

Im Controlling gehört der Umgang mit Daten zu den zentralen Aufgaben. Gleichzeitig steigen die Erwartungen an Aktualität, Genauigkeit und vorausschauende Steuerung [4]. In diesem Spannungsfeld rückt der Einsatz von KI in den Fokus. Die Technologie verspricht, bestehende Instrumente zu ergänzen und Entscheidungsprozesse datengetriebener zu gestalten. Allerdings ist unklar, wie gut die aktuellen Controlling-Strukturen auf diese Anforderungen vorbereitet sind.
Ein zentrales Problem besteht darin, dass viele Prozesse im Controlling auf historisch gewachsenen IT-Systemen und Berichtslogiken beruhen. Die Integration intelligenter Analyseverfahren erfordert daher nicht nur technische Anpassungen, sondern auch eine veränderte Sicht auf Informationsverarbeitung und Rollenverständnisse im Controlling. Herausforderungen bestehen in der Datenqualität sowie in der Transparenz und Nachvollziehbarkeit der eingesetzten Modelle. Hinzu kommt, dass viele Fachkräfte bislang nur begrenzte Erfahrung im Umgang mit KI sammeln konnten [3].
Darüber hinaus sind Verantwortlichkeiten im Controlling häufig arbeitsteilig organisiert. Der Einsatz lernender Systeme stellt diese Arbeitsteilung infrage, da Informationen nicht mehr nur gesammelt und berichtet, sondern auch interpretiert und gewichtet werden. Diese Entwicklung berührt nicht nur die operativen Abläufe, sondern auch das Rollenverständnis der Controlling-Funktion selbst [4] [3].
Eine Auseinandersetzung mit dem Status quo, mit möglichen Potenzialen und bestehenden Hürden kann helfen, aktuelle Entwicklungen einzuordnen – sowohl aus wissenschaftlicher Perspektive als auch mit Blick auf die betriebliche Praxis [3].

Zielsetzung

Vor dem Hintergrund der zunehmenden Verbreitung von Künstlicher Intelligenz (KI) in verschiedenen Unternehmensbereichen beschäftigt sich die vorliegende Arbeit mit der Relevanz und den Einsatzmöglichkeiten

von KI im Controlling. Ziel der Arbeit ist es, den aktuellen Status quo der KI-Nutzung in Controlling-Prozessen darzustellen und zukünftige Perspektiven zu analysieren. Dabei stehen sowohl die Chancen als auch die Herausforderungen im Fokus, die sich aus der Integration von KI in diesen zentralen Funktionsbereich ergeben.

Vorgehensweise

Die Arbeit ist in einen theoretischen und einen empirischen Teil gegliedert, wie in Abbildung 1 dargestellt.

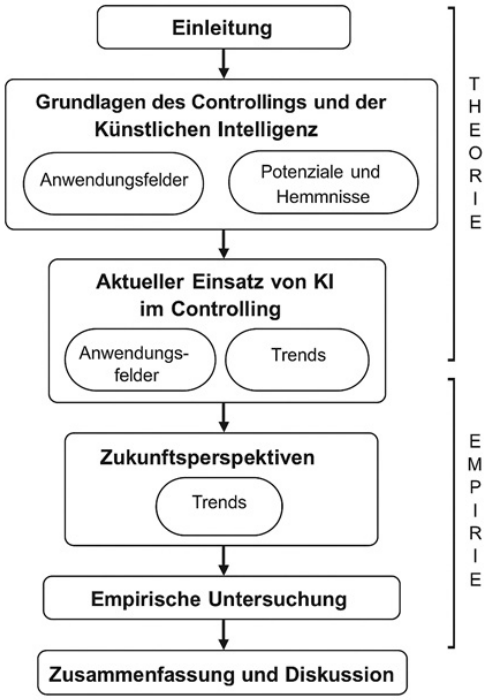


Abb. 1: Visualisierung des strukturellen Aufbaus der Arbeit mit theoretischem und empirischem Teil [2]

Der theoretische Teil legt die Grundlagen für das Verständnis der Thematik und erläutert zunächst die wesentlichen Aspekte des Controllings sowie die grundlegenden Konzepte der Künstlichen Intelligenz. Darauf aufbauend wird der aktuelle Einsatz von KI im Controlling untersucht. Dabei werden verschiedene Anwendungsfelder betrachtet und sowohl Potenziale als auch Hemmnisse analysiert. Kapitel 4 widmet sich den zukünftigen Entwicklungen. Hier werden auf Basis aktueller Studien und Fachliteratur Trends herausgearbeitet, die für den weiteren Einsatz von KI im Controlling von Bedeutung sein könnten.
Zur Ergänzung der theoretischen Betrachtung wird im empirischen Teil eine Umfrage unter Fach- und Führungskräften im Controlling durchgeführt. Der empirische Teil der Arbeit beginnt mit Kapitel 5. Ziel ist es, die aktuelle Nutzung von KI, bestehende Herausforderungen sowie Erwartungen an zukünftige Einsatzmöglichkeiten zu erfassen. Die Ergebnisse werden anschließend ausgewertet und mit den theoretischen Erkenntnissen in Beziehung gesetzt.

Ausblick

Der Einsatz von KI im Controlling steckt noch in den Anfängen. Erste Anwendungen zeigen Potenzial, doch viele Unternehmen zögern. Künftig wird es darauf ankommen, technische Lösungen sinnvoll in bestehende Prozesse zu integrieren und gleichzeitig die nötigen Kompetenzen im Umgang mit KI aufzubauen.
Auch das Rollenverständnis im Controlling wird sich weiterentwickeln. KI kann datengetriebene Entscheidungen unterstützen – ersetzt den Menschen aber nicht. Damit die Technologie ihr Potenzial entfalten kann, braucht es transparente Modelle, verlässliche Daten und eine klare strategische Einbindung.
Langfristig bietet KI nicht nur die Chance zur Effizienzsteigerung, sondern auch zur Neugestaltung klassischer Controlling-Aufgaben. Entscheidend wird sein, wie aktiv Unternehmen diesen Wandel mitgestalten.

Literatur und Abbildungen

[1] Statistisches Bundesamt. Etwa jedes achte Unternehmen nutzt künstliche Intelligenz. https://www.destatis.de/DE/Presse/Pressemitteilungen/2023/11/PD23_453_52911.html, 2023.

[2] Eigene Darstellung.

[3] Maximilian Feike, Bernd Bienzeisler, and Jens Neuhüttler. *Künstliche Intelligenz aus Sicht von Unternehmen*. Oliver Riedel, Katharina Hölzle, Wilhelm Bauer, 2024.

[4] Felix Niermann. Künstliche Intelligenz im Controlling: Potenziale erkennen und richtig nutzen. *CIC-Online*, 2023.

Large Language Models im Bereich Autonomes Fahren

Andreas Pitz **Markus Enzweiler**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung:

Spätestens seit der Veröffentlichung von ChatGPT Ende 2022 sind **Large Language Models** (kurz LLMs) sowie **Vision Language Models** (VLMs) der breiten Öffentlichkeit bekannt. Allerdings werden nur die wenigsten Nutzer dieser Modelle sich mit deren Architektur bzw. dem breiten Spektrum an Anwendungsmöglichkeiten auseinander gesetzt haben. LLMs und VLMs können weit mehr als nur zu chatten oder Texte zu generieren. Beispielsweise können moderne Modelle auch Bilder beschreiben oder Audio verarbeiten. Besonders ersteres kann gut für eine der großen technischen Herausforderungen unserer Zeit genutzt werden: Dem **Autonomen Fahren**, denn bei diesem ist die Kamera nach wie vor der wichtigste Sensor. Diese Arbeit ergründet den Nutzen von LLMs und VLMs in diesem Bereich und nutzt dafür als Basis den End-to-End Ansatz **LMDrive** von Shao et al. [3] basierend auf dem CARLA Simulator [1]. Bei CARLA handelt es sich um einen Open-Source Simulator zum Entwickeln, Trainieren und Validieren von Autonomen Systemen, basierend auf der Unreal Engine. Im Gegensatz zum modularen Ansatz mit seinen einzelnen Modulen für Wahrnehmung, Vorhersage und Trajektorienplannung, werden beim End-to-End Ansatz die Eingangsdaten der Sensoren durch ein neuronales Netz direkt in Kontrollbefehle umgewandelt. Ein Vorteil des End-to-End Ansatzes ist dabei die Fähigkeit der Modelle eigene Entscheidungen zu begründen und für Menschen verständlich auszudrücken, während einer der größten Nachteile im komplizierten debuggen der Systeme liegt. Durch die Möglichkeit eigene Entscheidungen zu begründen, kann z.B. das Vertrauen späterer Anwender gesteigert werden sowie das allgemeine Verständnis für die Vorgänge im Modell vergrößert werden.

LLM / VLM Aufbau und Fähigkeiten:

Der Unterschied zwischen LLMs und VLMs liegt in der Art von Inhalt den diese verarbeiten können. Während LLMs in der Lage sind in natürlicher Sprache (NLP / Natural Language Processing) zu interagieren, besitzen VLMs zusätzlich noch visuelle Fähigkeiten aus dem Bereich der Computer Vision. Sie enthalten also zusätzlich zu einem Sprach-Encoder noch einen Vision-Encoder, welcher in der Lage ist Farben, Formen, Texturen und weitere Bildmerkmale in Tokens umzuwandeln, die dann vom Modell verarbeitet werden können. Die Architektur moderner LLMs und VLMs basiert auf dem heute berühmten Paper *Attention is All you need* einer Gruppe von Google Forschern [4]. In diesem wird die *Transformer* Architektur vorgestellt, welche die Basis moderner Modelle dieser Art ist. Die Wichtigkeit des Transformers kann man auch an der ausgeschrieben Form von GPT - Generative Pretrained Transformer erkennen. Der Transformer ermöglicht sogenannten *Tokens*, numerischen, vektoriellen Darstellungen von natürlicher Sprache oder anderen Medien, einen reichhaltigen Kontext in sich aufzunehmen. Die Berechnungen im Transformer selbst sind dabei eine Reihe von Matrizenmultiplikationen mit lernbaren Gewichten bzw. Parametern. Moderne Modelle weisen heutzutage eine mehrstellige Milliarden Zahl an Gewichten auf. Diese Operationen im Transformer befähigen ein Modell nach dem Trainieren beispielsweise dazu, den Unterschied zwischen einer Parkbank und dem Geldinstitut Bank alleine anhand des umliegenden Kontexts zu erkennen. Durch diese erlernte Fähigkeit kann ein Modell auch differenzierte Aussagen zu Objekten im Straßenverkehr und deren Bedeutung für die Sicherheit in der gezeigten Umgebung liefern.

Vorgehensweise:

Im Zuge dieser Arbeit soll die Architektur des Ansatzes LMDrive um ein weiteres VLM zur Szenenerkennung erweitert werden. Dieses soll den genutzten Vision Encoder, welcher die Sensorinputs der Kamera verarbeitet, ersetzen. Die angepasste Architektur von LMDrive ist in Abbildung 1 zu sehen. Dabei werden im Zuge dieser Arbeit nur die Daten der Frontkameras genutzt und der LIDAR Sensor sowie die Rückkamera nicht weiter beachtet.

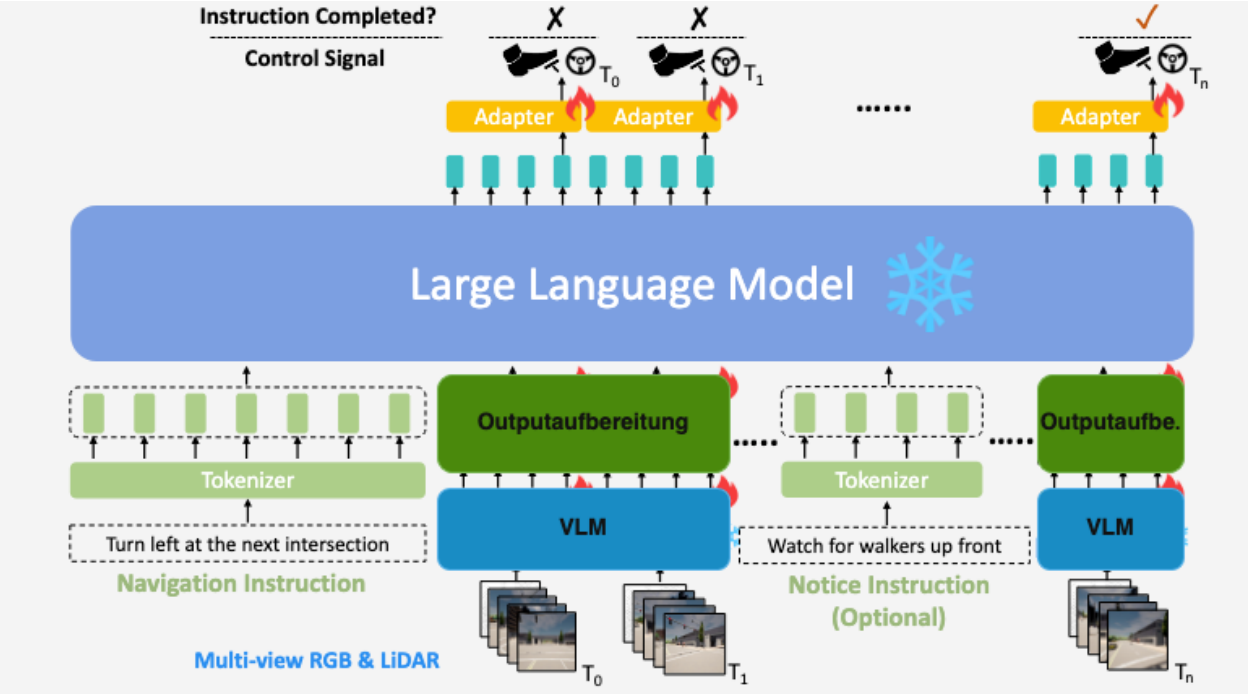


Abb. 1: LMDrive Architektur (angepasst) [3]

Als erster Schritt dieser Arbeit werden dafür verschiedene, in Frage kommende, Open-Source VLMs evaluiert und getestet um die passendsten Modelle für den Einsatz in die LMDrive Architektur zu finden. Dabei liegt der Fokus auf der Fähigkeit Bilder zu beschreiben und typische Objekte im Straßenverkehr auf diesen zu erkennen. Außerdem wird der *Prompt*, also die Aufgabenbeschreibung für die Modelle in natürlicher Sprache, optimiert. Das Ziel dabei ist es, hochwertige und vor allem konsistent formatierte Ergebnisse zu erhalten, die automatisiert weiter verarbeitet werden können. Abbildung 2 zeigt, dass moderne Modelle gute Ergebnisse in der Simulationsumgebung von CARLA erzielen können.

Abb. 2: Objekterkennung des Modells Qwen 2.5 VLM [2]

Ausblick:

Nach dem Evaluieren verschiedener Modelle anhand der oben genannten Kriterien auf Testdarstellungen, werden diese in die komplette Pipeline von LMDrive integriert. Die Effizienz der neuen Architektur wird anhand der gängigen Metriken und Benchmarks, welche in LMDrive genutzt werden, ausgewertet. Dazu zählt zum Beispiel die LangAuto Benchmark die Teil des Projekts LMDrive ist. Diese deckt alle wichtigen Verkehrssituationen und Konstellationen ab. Darunter zählen zum Beispiel Highways, Kreuzungen und Kreisverkehre. Außerdem enthält diese verschiedene Umwelt-, Tageszeit- und Wetterkonditionen um realitätsnahe Szenarios zu schaffen.

Literatur und Abbildungen

[1] Team CARLA. CARLA Simulator. https://carla.org, 01 2025.

[2] Hesham Mohamed Eraqi. CARLA Street Scene. https://www.researchgate.net/figure/Partial-road-blockages-added-to-CARLA-simulator-the-road-blockages-avoidance-algorithm_fig1_364776163, 2022.

[3] Hao Shao et al. LMDrive. https://arxiv.org/abs/2312.07488, 12 2023.

[4] Ashish Vaswani et al. Attention is All you need. https://arxiv.org/abs/1706.03762, 06 2017.

Mapping Customer Feedback to Roadside Assistance Customer Journey Steps: A Proof of Concept Using Supervised Machine Learning

Luca Raichle **Steffen Schober**

Department of Computer Science and Engineering, Esslingen University

Work carried out at Mercedes-Benz AG, Stuttgart-Vaihingen

Introduction

Customer feedback is being considered as one of the most valuable sources of insight within the aftersales departments of large companies. However, leveraging this feedback effectively still remains a challenge, primarily due to the vast amount of unstructured textual data. Manually reviewing millions of feedback is inefficient in many regards, especially in terms of time, scalability and consistency. [1] Recent machine learning improvements have made it easier for companies to extract structured insights from large-scale unstructured data, allowing them to work more effectively with customer feedback and make data-driven decisions. This opens up new opportunities to gain deeper insights and take a step closer to the ultimate goal in aftersales: increasing customer satisfaction.

Motivation and Background

At Mercedes-Benz, millions of customer feedback entries are collected through a CRM system. The Business Excellence Acceleration Team (BEAT) took on the challenge of addressing the unstructured feedback dilemma using machine learning techniques, initially using a topic modelling approach. Within this approach, key concerns were extracted from the customer feedback to identify general pain points. This modern approach helped to gain deeper insights into customer problems, but over time it became clear, that the extracted topics were often too generic in specific use cases, making it hard to filter common problems and address them at their root causes.
Based on the limitations of the initial topic modeling approach, a new direction was introduced - trying to identify more specific customer issues within a more structured context. As part of a broader Mercedes-Benz initiative, customer journeys were defined to outline the interactions of a customer with Mercedes-Benz after the initial purchase of the vehicle. This opened a new opportunity to combine two key areas in aftersales: customer feedback and customer journeys.

Goal of the Thesis

The goal of the Bachelor Thesis is to develop a proof of concept for mapping customer feedback onto specific steps of a customer journey using machine learning classification techniques. Taking the already defined Roadside Assistance Customer Journey (RSA) as a starting point, the concept of using classification to extract deeper insights in a specific customer journey step shall be proven in this specific use case. For each journey step, the objective is to present the specific customer feedback case descriptions, the number of cases per step, the top topics within the step and the overall sentiment. Based on this successful mapping, relevant stakeholders shall be able to work closer with customer feedback.
Because of the reason that there are multiple customer journeys within Mercedes-Benz, this proof of concept remarks the starting point for a potential series of projects, which feasibility, validity and best approach shall be verified with this Bachelor Thesis - ultimately providing a framework that is scalable to other customer journeys across the organization.

Methodological Approach

This Bachelor Thesis follows the widely used standard for machine learning projects called "Cross-Industrial Standard Process and Data Mining" (CRISP DM). There, six key phases are defined with specific objectives and outcomes after each phase. Based on these six phases, the relevant work items were identified and scheduled accordingly to ensure a smooth and structured project flow.

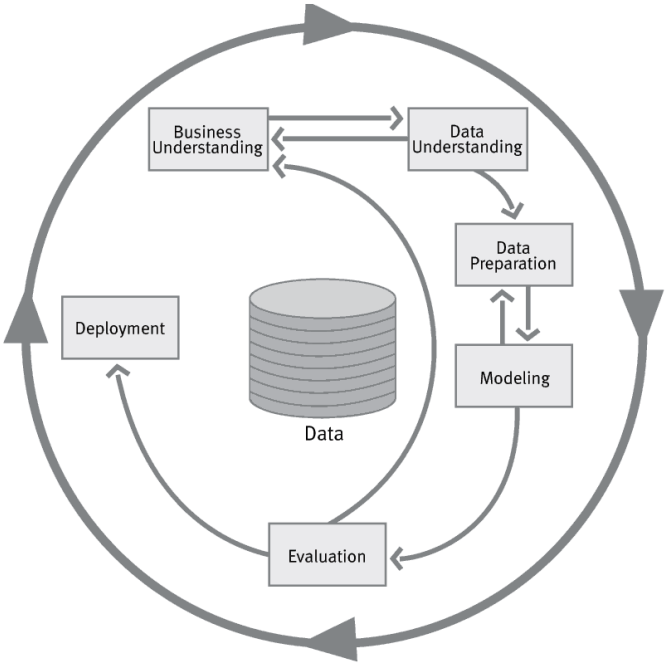


Fig. 1: 6 Phases of the CRISP-DM Framework [4]

The first phase of the CRISP-DM framework called "Business Understanding" focuses on understanding the key problem that shall be solved within the project. This includes assessing the initial situation, determining objectives and goals, producing a project plan and gathering and understanding the key requirements from relevant stakeholders of the project. [3] In the context of the Customer Journey Mapping project, workshops and several interviews were conducted to ensure the alignment of the project scope with the interests of the two cooperating departments. Furthermore, the current state and its pain points were examined to confirm the validity of the project while the potential future state with its opportunities was analyzed.
In the second phase, "Data Understanding", the main goal is to analyze the available data sources and assess them regarding quantity and quality. This phase also includes creating the relevant dataset for the project and visualizing the data. [3]

Fig. 2: Creation of the Thesis Data Set [2]

To create the dataset, two data sources were combined to include all relevant information. In the next step, the dataset was filtered to include only Roadside Assistance cases from the year 2024 that contain a non-empty description. A key challenge in the process of the EDA was to work with unstructured data, which made it difficult to apply traditional EDA metrics. Therefore, the focus was more on identifying null values, filtering out non-relevant columns and dive into some specific insights such as average feedback length and more.
The key objective of the third phase, "Data Preparation", is to prepare the data for the modeling phase. [3] In the context of the Customer Journey Mapping project, this involves initially labeling a sample set of data, that can be used to train the model. This initial data labeling process would involve a significant amount of manual effort, requiring thousands of cases to be labeled by hand in order to generate sufficient training data for the model. Therefore, the initial labels shall be generated using few-shot prompting with a Large Language Model and journey step descriptions to minimize manual efforts. Based on the validation of the results, it may be necessary to further optimize the LLM labeling approach.
The fourth phase called "Modeling", aims to train a supervised classification model based on the previously labeled training data to be able to classify further customer feedback. [3] Despite the impressive performance of LLMs like GPT or Gemma on few-shot tasks such as initial labeling, their high cost, slow data processing times and resulting limited scalability make them less suitable for the Customer Journey Mapping use case. Therefore, based on the initial labeling, a custom machine learning model will be developed aiming to reach the performance of the LLMs while significantly improving efficiency and

reducing computational costs. Training a separate model is necessary due to the large volume of cases - involving hundreds of thousands of cases in this proof of concept, with the potential to scale to millions in future projects.
The fifth phase of the CRISP-DM Framework called "Evaluation" deals with the review of the model results. [3] There, machine learning metrics such as accuracy, F-1 score and confusion matrix will be used to evaluate the performance of the trained models. Furthermore, manual reviews of the labelling shall be used in order to review its performance with expertise.
After successfully completing the phases above, the model can be deployed.

Expected Outcome and Next Steps

The expected outcome of this project is the successful completion of the proof of concept demonstrating the performance and feasibility of mapping customer feedback onto Customer Journey steps using supervised machine learning to extract actionable insights. The project also aims to evaluate, if this approach and methodology can be applied to other Customer Journeys. Furthermore, the results will be shared with relevant stakeholders of the journey steps - bringing customer feedback directly back to the people in charge and enabling them to proactively improve products and services.

References and figures

[1] Evert De Haan, Manjunath Padigar, Siham El Kihal, Raoul Kübler, and Jaap Wieringa. Unstructured data research in business: Toward a structured approach. https://www.sciencedirect.com/science/article/pii/S0148296324001590, 2024.

[2] Own representation.

[3] Andre Shimaoka, Alfredo Goldman, and Renato Ferreira. The evolution of CRISP-DM for Data Science: Methods, Processes and Frameworks. https://www.researchgate.net/profile/Renato-Ferreira-22/publication/384999724_The_Evolution_of_CRISP-DM_for_Data_Science_Methods_Processes_and_Frameworks/links/6710e3de09ba2d0c760588e2/The-Evolution-of-CRISP-DM-for-Data-Science-Methods-Processes-and-Frame, 2024.

[4] Laurenz Wuttke. CRISP-DM: Grundlagen, Ziele und die 6 Phasen des Data Mining Prozess. https://datasolut.com/crisp-dm-standard/, 2023.

Einsatz von Low-Code/No-Code-Plattformen zur KI-basierten Prozessautomatisierung: Entwicklung und Evaluation eines Prototyps

Tom Rehm **Manfred Schoch**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

Die digitale Transformation und der Vormarsch Künstlicher Intelligenz (KI) stellen Unternehmen vor die Herausforderung, ihre Geschäftsmodelle agil und zukunftssicher zu gestalten. Softwareentwicklung ist dabei ein Schlüsselfaktor, jedoch kämpfen viele Betriebe mit dem Mangel an IT-Experten und hohen Entwicklungskosten. Low-Code/No-Code (LC/NC)-Plattformen versprechen hier Abhilfe, indem sie die Anwendungsentwicklung auch für sogenannte Citizen Developer ohne tiefgreifende Programmierkenntnisse zugänglich machen [6]. Gleichzeitig eröffnen moderne KI-Modelle, insbesondere Large Language Models (LLMs), neue Potenziale für die Automatisierung komplexer Aufgaben.

Zielsetzung

Für die Arbeit wurde bei einem großen Automobilbauer das Qualitätsmanagement befragt und eine Reihe zeitaufwendiger und repetitiver Aufgaben identifiziert, die ein hohes Automatisierungspotenzial durch KI-Agenten aufweisen. Dazu zählen die Optimierung der Fehler-Ticket-Bearbeitung, intelligentes Postfachmanagement, Unterstützung bei der Testfallerstellung, zentraler Wissenszugriff sowie die Automatisierung des Reportings. Die zentrale Problemstellung lag darin, zu evaluieren, wie LC/NC-Plattformen genutzt werden können, um KI-gestützte Automatisierungslösungen für solche QM-Prozesse effizient zu konzipieren und prototypisch umzusetzen. Um daraus ein Fazit zu ziehen, ob ein solches Vorgehen einen praktischen Nutzen hat.

Umsetzung

Nach einer sorgfältigen Analyse der Anforderungen, In der Form einer Befragung und dem Vergleich verschiedener Entwicklungsansätze, darunter die Programmierung mit Frameworks wie LangChain als Gegensatz zu Low-Code-/No-Code-Plattformen, fiel die Entscheidung auf die Open-Source-Plattform n8n. Diese Plattform stellte einen geeigneten Kompromiss zwischen Flexibilität, insbesondere durch die Möglichkeit der Nutzung von Code-Nodes und API-Integrationen, sowie der Option des Self-Hostings dar, was insbesondere im Hinblick auf Kostenkontrolle und Datensicherheit von Bedeutung war. Zudem überzeugte n8n durch eine aktive Community, die eine nachhaltige Weiterentwicklung unterstützt.

Im Rahmen der prototypischen Umsetzung kamen mehrere zentrale Technologien zum Einsatz. Die Workflow-Orchestrierung und zentrale Integrationsschicht wurde durch n8n als Low-Code-/No-Code-Plattform realisiert. Für komplexe Sprachverarbeitungsaufgaben diente GPT-4o von OpenAI als zentrales Element („Gehirn“) der Agenten. Ergänzend wurden kleinere, spezialisierte Modelle, wie etwa das Google FlashModel, für repetitive und weniger anspruchsvolle Aufgaben in Betracht gezogen. Zur semantischen Suche wurde das Embedding-Modell text-embedding-3-small von OpenAI verwendet, um Text-Chunks zu vektorisieren. Dabei werden Textbausteine auf Grundlage ihrer Eigenschaften in einen Vektor der Länge 1536 transformiert, vergleichbar mit der vereinfachten Darstellung in Abbildung 1, bei der zur Veranschaulichung lediglich zwei Dimensionen berücksichtigt werden. Die resultierenden Vektordaten wurden in einer Supabase-Datenbank mit der *pgvector*-Erweiterung für PostgreSQL gespeichert und effizient abgefragt, was insbesondere für Retrieval-Augmented Generation (RAG) von zentraler Bedeutung war. Dies stellte das Kernelement des in Abbildung 2 gezeigten Prozesses dar, der die Interaktion mit in der Cloud gespeicherten Dateninhalten über ein Sprachmodell ermöglicht.

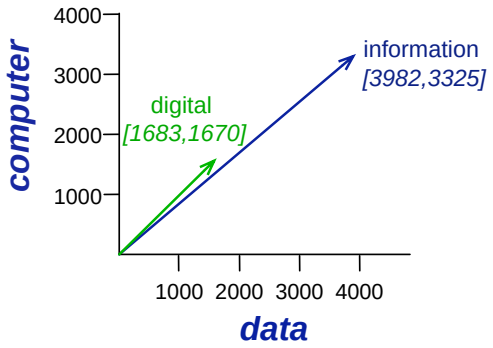


Abb. 1: Veranschaulichung von Vektor-Embeddings anhand von zwei Dimensionen [3]

Das von Anthropic Ende 2024 eingeführte Model Context Protocol (MCP) [5] ermöglicht eine verbesserte Anbindung von KI-Agenten an externe Systeme, wodurch ein umfassender Zugriff auf Jira-Einträge in Form strukturierter Testfälle umgesetzt werden konnte..

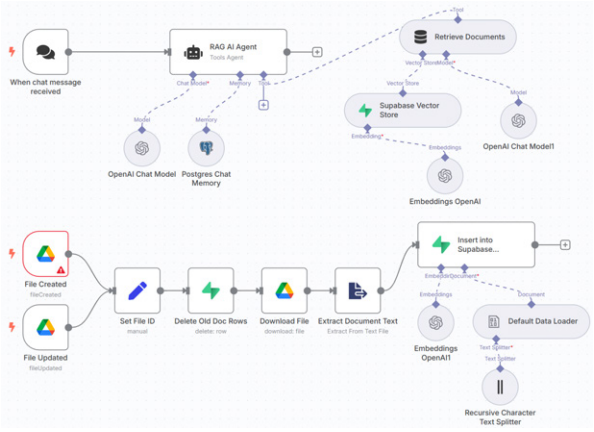


Abb. 2: Workflow in n8n zur Dokumentensuche [1]

Risiken

Durch diese Art der Automatisierung entstehen jedoch auch Risiken. Low-Code-Plattformen bergen ein erhöhtes Risiko eines sogenannten *Vendor Lock*-ins, da sie meist auf proprietäre Modellierungswerkzeuge, Datenformate und Logiken setzen und viele ausschließlich in der Cloud verfügbar sind. Die Kosten sind oft nutzer- oder transaktionsbasiert, was Skaleneffekte bei steigender Nutzerzahl einschränkt [2].
Darüber hinaus birgt der Einsatz von KI-Modellen, die das Zentrum der KI-Automatisierung darstellen, das Risiko, dass es bei der Ausführung von Workflows zu unentdeckten Fehlern kommt, verursacht durch typische KI-Probleme wie beispielsweise Halluzinationen. Sogenannte *Prompt-Injection*-Angriffe stellen ein weiteres Risiko dar, weshalb besonders darauf geachtet werden sollte, auf welche Ressourcen der jeweilige Agent Zugriff hat – insbesondere im Hinblick auf angebundene APIs [4].
Zudem begünstigt die benutzerfreundliche Handhabung maßgeblich das Risiko, dass sich eine sogenannte Schatten-IT entwickelt.

Ausblick

Low-Code-Umgebungen könnten künftig durch zunehmend leistungsfähige KI-Systeme ersetzt werden, da diese immer weniger menschliche Einflussnahme erfordern. Gleichzeitig erscheint jedoch auch eine Art Symbiose denkbar: Der von KI generierte klassische Code bleibt häufig nur für Mitarbeitende mit fundiertem IT-Hintergrund verständlich. Low-Code-Ansätze könnten daher künftig eine wichtige Rolle dabei spielen, komplexe KI-Workflows zugänglicher und nachvollziehbarer zu machen. Ein vollständiger Ersatz klassischer Programmierung durch Low-Code ist jedoch eher unwahrscheinlich.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Edona Elshan and Björn Binzer. Mehr als ein Trend?: Wie Low-Code die digitale Transformation unterstützt. *HMD Praxis der Wirtschaftsinformatik*, 2024.

[3] Daniel Jurafsky and James H. Martin. *Vector Semantics and Embeddings*. Stanford University, 2024.

[4] Matthew Kosinski. What is a prompt injection attack? https://www.ibm.com/think/topics/prompt-injection, 2024.

[5] Anthropic PBC. Introducing the Model Context Protocol. https://www.anthropic.com/news/model-context-protocol, 2024.

[6] Apurvanand Sahay, Arsene Indamutsa, Davide Di Ruscio, and Alfonso Pierantonio. *Supporting the understanding and comparison of low-code development platforms*. IEEE, 2020.

Code-Generierung durch Open Source LLMs

Zisis Relas **Mirko Sonntag**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einführung

Der Begriff „Vibe Coding" gewinnt derzeit zunehmend an Beliebtheit. Darunter wird verstanden, dass AI-Tools wie ChatGPT, Cursor oder Copilot zunehmend die Coding-Arbeit von Entwicklern abnimmt. Hierbei wird der Software-Code durch natürliche Sprache generiert. [3] Die zunehmende Beliebtheit unterstreicht die wachsende Bedeutung die Large Language Models (LLMs) im Bereich der Softwareentwicklung einnehmen. Die derzeit populären AI-Tools basieren in der Regel auf leistungsstarken, cloud-basierten. proprietären Modellen die über eine hohe Parameteranzahl und großen Kontextfenstern verfügen. Das führt zu beeindruckenden Ergebnissen in der Generierung von Code für die Nutzer dieser Systeme. Gleichzeitig gehen solche Lösungen aber auch mit Nachteilen im Bereich des Datenschutzes und der Anpassbarkeit einher. Open-Source-LLMs, die lokal auf eigener Hardware betrieben werden können, stellen in diesem Zusammenhang eine vielversprechende Alternative dar. Sie ermöglichen ein hohes Maß an Kontrolle über das Modellverhalten und lassen sich gezielt für domänenspezifische Anwendungen anpassen.

Zielsetzung der Arbeit

Die begrenzten Rechenressourcen mit lokaler Hardware stellen bei der Code-Generierung mit Open-Source-LLMs eine besondere Herausforderung dar. Vor diesem Hintergrund stellt sich die Frage: Wie kann mithilfe von Finetuning des Modells und Prompt-Engineering-Techniken die Qualität des Code-Outputs verbessert werden im Bereich der Webentwicklung? Zur Beantwortung der Frage wird im Rahmen der Arbeit eine Softwarelösung als Erweiterung für die Open-Source-Entwicklungsumgebung Visual Studio Code entwickelt. Die Erweiterung soll die Generierung von lauffähigen Software-Code auf Basis natürlicher Spracheingaben ermöglichen und daraus resultierend die erforderlichen Dateien automatisiert erstellen. Die der Lösung zugrunde liegenden Modelle sollen Open-Source sein und lokal betrieben werden.

VS Code Extension Development

Visual Studio Code ist eine von Microsoft entwickelte integrierte Open-Source-Entwicklungsumgebung (IDE). Aufgrund seiner modularen Architektur bietet VS Code eine umfangreiche Extension-API an, womit zahlreiche Aspekte der Benutzeroberfläche und Funktionalität erweitert, werden können. Im Rahmen der Arbeit wird eine Erweiterung entwickelt, die über Webviews individuelle Funktionalitäten bereitstellt, die über den Umfang der VS Code API hinausgehen. Die Webview-Komponente dient als graphisches User Interface zur Interaktion mit dem LLM. Die Auswahl des LLMs, die Eingabe von natürlichen Sprachbefehlen (Prompts), die Anzeige des generierten Codes sowie die anschließende Erzeugung und Persistierung des Quellcodes wird durch das GUI ermöglicht.

Hosting der LLMs

Zur lokalen Ausführung von LLMs wird in dieser Arbeit die Open-Source-Anwendung Ollama verwendet. Ollama ermöglicht die Ausführung von diversen LLMs auf der eigenen Hardware und stellt hierfür sowohl eine Kommandozeilenschnittstelle (CLI) als auch eine Programmierschnittstelle (API) zur Verfügung, über die eine direkte Interaktion mit den Modellen erfolgen kann. Ein wesentlicher Aspekt von Ollama ist die Unterstützung von Finetuning durch LoRA (Low-Rank Adaptation), das über sogenannte Modelfiles realisiert wird. Dies erlaubt eine gezielte Anpassung von LLMs an spezifische Anwendungsbereiche oder Domänen. [1]

Anwendungsüberblick

Die Visual-Studio-Code-Erweiterung kann über einen Befehl innerhalb der IDE gestartet werden. Beim Aufruf öffnet sich ein neues Editor-Fenster, das die graphische Benutzeroberfläche (GUI) bereitstellt. Über die GUI lassen sich spezifische Anforderungen wie die Technologieauswahl oder die funktionalen Anforderungen an den zu generierenden Code definieren. Das LLM generiert basierend auf den definierten Anforderungen im Prompt

den Softwarecode und zeigt ihn in einer Editor-Ansicht im Nachrichtenverlauf an. Der generierte Code kann anschließend manuell weiterbearbeitet oder durch zusätzliche Prompts iterativ optimiert werden. Anschließend lässt sich der finale Code in Dateien überführen und lokal persistieren. Ein schematischer Überblick der Softwarearchitektur ist in Abbildung 1 dargestellt.

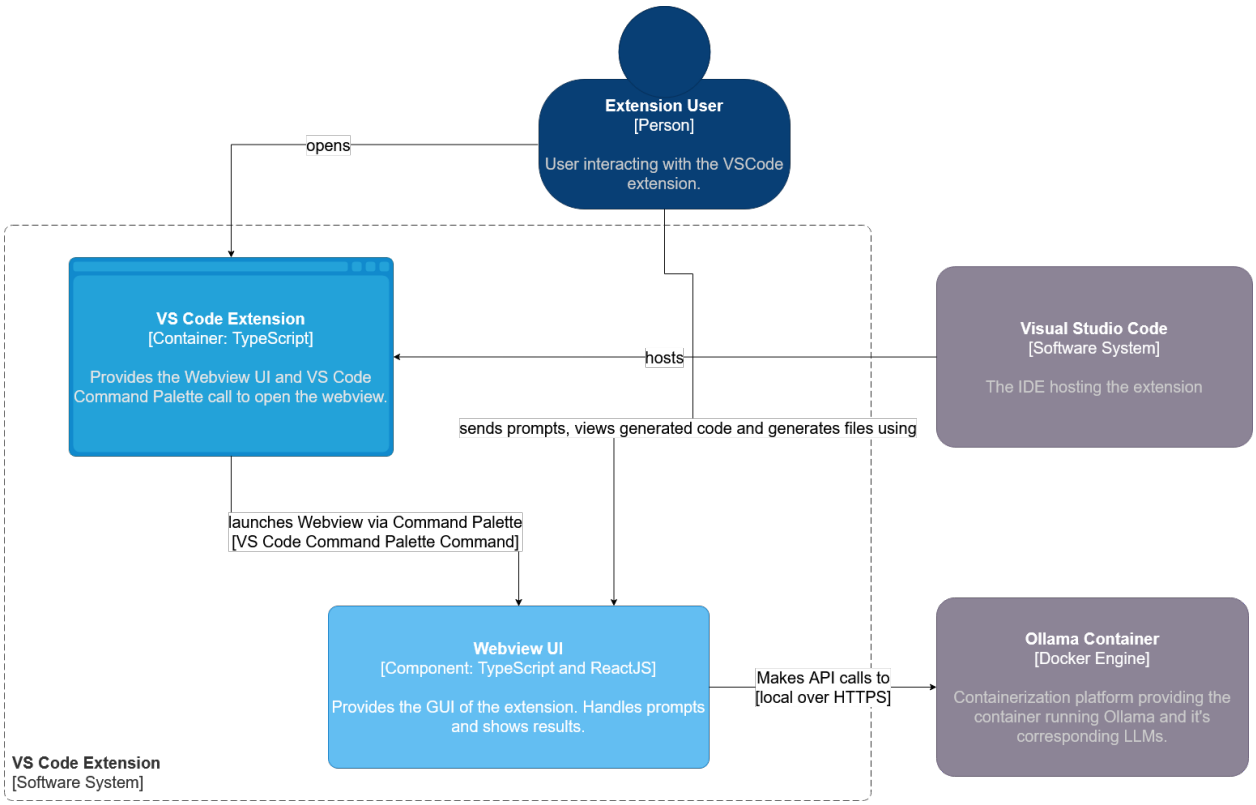


Abb. 1: C4-Modell Container-Diagramm [2]

Optimierung der Code-Qualität

Der initiale Ansatz zur Codegenerierung basiert auf Zero-Shot-Prompting. Bei diesem Ansatz wird das LLM ohne zuvor gezeigte Beispiele direkt zur Generierung von Softwarecode aufgefordert. Die Qualität des erzeugten Codes hängt maßgeblich vom domänenspezifischen Pre-Training vom LLM ab. Im weiteren Verlauf wird untersucht, ob durch den Einsatz verschiedener Prompting-Techniken die Qualität des generierten Codes im Hinblick der Erfüllung der Anforderungen steigern lässt. Auch die Auswahl des LLMs kann hierbei einen zentralen Faktor darstellen. So zeigen Reasoning-Modelle wie Deepseek R1 eine besondere Eignung für komplexe Softwareanforderungen, da sie in der Lage sind, problembezogene Anforderungen in logische Einzelschritte zu zerlegen und diese systematisch abzuarbeiten. [4] Ein weiterer Ansatz zur Qualitätssteigerung ist das Finetuning vom LLM. Durch das Beschränken vom Anwendungskontext, wie den genutzten Technologien und Frameworks, und domänenspezifischen Nachtraining kann die Ergebnisqualität verbessert werden.

Erste Ergebnisse und Ausblick

Die Anwendung generiert mit vortrainierten LLMs (Pre-Trained LLMs) in der Regel Softwarecode, der die im Prompt formulierten Anforderungen zumindest teilweise erfüllt. Dabei variiert die Qualität der Ergebnisse jedoch deutlich, abhängig vom eingesetzten Modell, der verwendeten Prompting-Technik sowie der Komplexität der Anforderungen. Insbesondere das Reasoning-Modell Deepseek R1 zeigt bei komplexeren Full-Stack-Anwendungen bessere Ergebnisse als General Purpose Models wie CodeLlama oder Granite-Code. Gleichzeitig lässt sich feststellen, dass die begrenzte Kontextgröße (Context Size) lokaler Modelle zu einer geringeren Konsistenz und Qualität der generierten Ergebnisse führt als bei kommerziellen Cloud-basierten Lösungen wie den OpenAI-Modellen. Für die weitere Arbeit ist geplant diese Limitierungen durch gezielte Optimierungen beim Prompting und Finetuning zu adressieren.

Literatur und Abbildungen

[1] Tahir Balarabe. What is Ollama: Running Large Language Models Locally. https://medium.com/@tahirbalarabe2/what-is-ollama-running-large-language-models-locally-e917ca40defe, 03 2025.

[2] Eigene Darstellung.

[3] Shalini Harkar. What is vibe coding? https://www.ibm.com/think/topics/vibe-coding, 04 2025.

[4] Aili McConnon. DeepSeek's reasoning AI shows power of small models, efficiently trained. https://www.ibm.com/think/news/deepseek-r1-ai, 01 2025.

Stand und Herausforderungen beim Einsatz von KI-gestützten Empfehlungssystemen im digitalen Marketing

Valentina Resch **Anke Bez**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

Künstliche Intelligenz (KI) ist längst mehr als ein abstraktes Forschungsfeld – sie prägt unseren Alltag und verändert zunehmend auch die Art und Weise, wie Unternehmen mit ihren Kunden kommunizieren. Besonders im digitalen Marketing ermöglichen KI-basierte Empfehlungssysteme eine gezielte, datengestützte Nutzeransprache – etwa durch personalisierte Produktempfehlungen, Inhalte oder Newsletter. Die Verknüpfung von KI und Marketing schafft neue Formen der Interaktion, bringt aber zugleich technische, ethische und wirtschaftliche Herausforderungen mit sich.

Zielsetzung und methodisches Vorgehen

Ziel dieser Arbeit ist es, den aktuellen Entwicklungsstand von KI-gestützten Empfehlungssystemen im digitalen Marketing zu analysieren. Im Fokus stehen zentrale Fragestellungen zur technischen Umsetzung, zur Nutzerakzeptanz und zu relevanten Erfolgsfaktoren in der Praxis. Neben der theoretischen Analyse von Grundlagen, Funktionsweisen und Anwendungsfeldern ist eine empirische Untersuchung geplant, die durch Umfragen und Interviews ergänzt wird. KI-Empfehlungssysteme im Marketingkontext – aktueller Stand und Herausforderungen Bereits die theoretische Auseinandersetzung zeigt: Die technische Entwicklung im Bereich KI-gestützter Empfehlungssysteme vollzieht sich in einem immer schneller werdenden Tempo – ihre unternehmerische Umsetzung hingegen bleibt vielfach hinter den vorhandenen Möglichkeiten zurück. Besonders im internationalen Vergleich wird deutlich, wie stark die Implementierung variiert: Während Unternehmen in den USA oder China KI-basierte Empfehlungssysteme längst als festen Bestandteil datengetriebener Marketingstrategien einsetzen – etwa durch personalisierte Produktvorschläge bei Amazon oder gezielte Inhaltsausspielung bei Alibaba –, zeigen sich viele deutsche, insbesondere mittelständische Unternehmen noch zurückhaltend [5]. Ein wesentlicher Grund dafür liegt in der hohen Komplexität der Einführung solcher Systeme. Neben technologischer Kompetenz bedarf es einer verlässlichen Datenbasis, strategischer Planung und organisatorischer Offenheit für Veränderung [3]. Gerade in kleinen und mittleren Unternehmen fehlt es häufig an diesen Voraussetzungen, was die Umsetzung trotz vorhandener technischer Möglichkeiten erschwert. Dabei reicht es nicht aus, klassische Verfahren wie kollaboratives oder inhaltsbasiertes Filtern einfach durch KI-gestützte Methoden wie maschinelles Lernen, semantische Verarbeitung oder neuronale Netze zu ersetzen [6]. Entscheidend ist vielmehr eine durchdachte Einbettung dieser Systeme in bestehende Prozesse sowie eine klare Zieldefinition. Zusätzlich zu den technischen und strukturellen Herausforderungen treten auch ethische Fragen zunehmend in den Vordergrund. Dazu gehören unter anderem die mangelnde Transparenz algorithmischer Entscheidungen, mögliche Verzerrungen in den Empfehlungen sowie Unsicherheiten hinsichtlich der Akzeptanz durch Nutzende. Trotz dieser Herausforderungen zeigen erfolgreiche Praxisbeispiele, dass durch gezielte Segmentierung und personalisierte Ansprache wirtschaftliche Erfolge erzielt werden können – etwa in Form verbesserter Conversion Rates oder nachhaltigerer Kundenbindung [4].

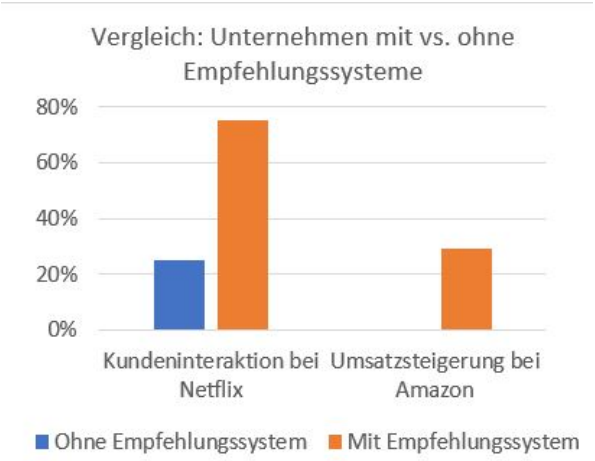


Abb. 1: Vergleich von Kundeninteraktion und Umsatz mit und ohne Empfehlungssystem [1]

Fazit und Ausblick

KI-gestützte Empfehlungssysteme werden künftig noch stärker individualisierte und kontextabhängige Vorschläge generieren – auch durch Fortschritte in generativer KI und multimodaler Datenverarbeitung. Gleichzeitig wird die Notwendigkeit wachsen, Transparenz und Datenschutz in der Kundeninteraktion stärker zu berücksichtigen. In Deutschland sind hier Regulierungen wie die DSGVO wegweisend, international entwickeln sich jedoch unterschiedliche Rahmenbedingungen [2]. Unternehmen, die es schaffen, KI verantwortungsbewusst, strategisch sinnvoll und nutzerzentriert einzusetzen, werden im digitalen Marketing der Zukunft einen deutlichen Wettbewerbsvorteil erzielen.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Generaldirektion Kommunikationsnetze European Union. Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). https://digital-strategy.ec.europa.eu/de/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence, 04 2021.

[3] André Klahold. *Empfehlungssysteme*. Springer Gabler, 2013.

[4] Stefan Ponitz. Chancen und Risiken beim Einsatz von Künstlicher Intelligenz in Unternehmen. https://www.fokus-ki.de/ki-strategie/chancen-und-risiken-beim-einsatz-von-kunstlicher-intelligenz-in-unternehmen/, 11 2024.

[5] 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, 10 2024.

[6] Meike Terstiege. *KI in Marketing & Sales – Erfolgsmodelle aus Forschung und Praxis*. Springer Gabler, 2022.

Begleitende Evaluation der SAP Global Trade Service Umstellung bei Andritz Schuler

Frederik Riesel **Thomas Rodach**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Andritz Schuler, Göppingen

Einleitung

Am Ende eines abgeschlossenen Projektes steht immer eine Evaluation an, diese Sorgt dafür, dass dieselben Fehler in den darauffolgenden Projekten nicht wiederholt werden. Dafür muss allerdings schon während der Durchführung des Projektes darauf geachtet werden, die vermeidbaren Fehler zu erkennen und zu Protokoll zu bringen. In diesem Artikel geht es um ein Projekt zur Umstellung des SAP Global Trade Services. Hierbei handelt es sich um die Software zur Zollabwicklung und Sanktionsprüfung. Diese Bereiche sind essenziell wichtig, da ohne eine solche Software ein Unternehmen keine Handelsbeziehungen zu Drittländern erhalten kann. Die Sanktionsprüfung erfolgt, um auszuschließen, dass an Länder Waren oder Dienstleistungen versendet werden. Im Allgemeinen handelt es sich um Embargos auf bestimmte Produkte, wie Waffen oder Munition, es kann allerdings auch vorkommen das die gesamte Handelsbeziehung mit ausgewählten Ländern gestoppt werden. Diese Art von Sanktionen erfolgt zum Beispiel gegen Länder wie den Iran oder Nordkorea.

Ziel der Arbeit

Die Bachelorthesis gibt einen Überblick über die gängigen Projektmanagementmethoden, die heutzutage bei Softwareprojekten in Unternehmen eingesetzt werden. Zudem geht es um die Evaluation eines Softwareprojektes. Es werden mithilfe eines Fragebogens die Projektteilnehmern befragt, wie ihrer Meinung nach das Projekt bisher abläuft und was die Teilnehmer bereits für Verbesserungsvorschläge für die Projektleitung haben. Bei dem Projekt geht es konkret um ein Umstellungsprojekt, der Zollabwicklung und Sanktionsprüfung. Aus verschiedenen Gründen hat die Firma sich dafür entschieden, den bisherigen Vertrag mit dem Anbieter zu kündigen und auf eine SAP-eigene Lösung umzustellen.

Wofür werden Finanzsanktionsprüfungen benötigt?

Die Finanzsanktionsprüfungen stellen sicher, dass keine Waren, Dienstleistungen oder Gehälter, an Personen oder Unternehmen, verkauft und ausgezahlt werden, die auf einer, von der EU erstellten, Liste stehen. Diese Finanzsanktionslisten werden ständig aktualisiert. Des Weiteren müssen diese Listen gekauft und bezahlt werden. Auf diesen Listen stehen verschiedene Staaten wie bereits erwähnt, aber eben auch Unternehmen oder sogar Personen. [5] Als Beispiel kann man den FC Chelsea anführen. Auf den ersten Blick ein englischer Fußballclub, aber wenn man sich mit dessen Unternehmensstruktur auseinandersetzt, wird klar, dass der Eigentümer des Clubs auf den Finanzsanktionslisten der EU aufgeführt wird. Aus diesem Grund darf an den FC Chelsea keine Waren oder Dienstleistungen verkauft werden.

Zollabwicklung

Bei der Zollabwicklung geht es um für den Export bestimmte Waren, die eine Ausfuhranmeldung benötigen. Der Zoll muss die Waren, vor dem Übertritt der Grenze, kontrollieren und die dafür anstehenden Zölle, die auf importierte und exportierte Waren erhoben werden, einnehmen. [6] Bei dieser Kontrolle wird auch festgestellt, ob die Waren auch an den vorgesehenen Empfänger geliefert werden dürfen. Innerhalb der EU und auch Staaten wie die Schweiz, die nicht der Europäischen Union beigetreten sind, aber dem Binnenmarktabkommen zugestimmt haben, können Waren auch ohne Zollprüfung leichter und mit weniger verbundenen Kosten gehandelt werden.

SAP-Einführungsprojekte

Bei Einführungsprojekten im SAP-Umfeld ist die Vorarbeit nicht nur sehr wichtig, sondern sogar zwingend notwendig. Zum einen werden Altlasten

abgeworfen und so nicht in das neue System übernommen. Zu sogenannten Altlasten gehören zum Beispiel Customizing-Programme, die schon länger nicht mehr genutzt werden oder mittlerweile ersetzt worden sind. [3] Zum anderen ist die Umstellung mit einem aufgeräumten System deutlich einfacher. Somit sind Punkte wie die Stammdatenpflege wichtig zu klären, die Stammdaten sollten auf dem neuesten Stand sein und möglichst alle Produkte des Unternehmens umfassen, aber ebenfalls von Altlasten befreit sein, bevor das Projekt an den Start geht. Bei solchen Umstellungsprojekten müssen vorher wichtige Fragen beantwortet werden, wie zum Beispiel, sind die betroffenen Prozesse auf dem neusten Stand und in einer entsprechenden Software detailliert abgebildet. Und ebenfalls sehr wichtig sind die Verantwortlichkeiten innerhalb des Umstellungsprojektes vollständig geklärt und allen Teilnehmern klar. Also ist die Hierarchie allen Projektteilnehmern bekannt. Diese Schritte sind wichtig, um die Abwicklung der Umstellung so einfach wie möglich zu halten. Da die technische Umsetzung dieser schon für sich genommen sehr kompliziert ist. Das gesamte Umstellungsprojekt ist dann noch komplex und benötigen genaue Planung und Steuerung. Deshalb sind genau vordefinierte Verantwortlichkeiten essenziell wichtig. [2]

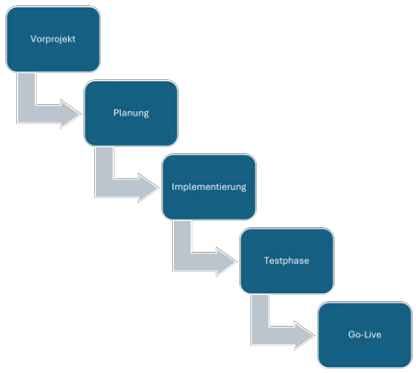
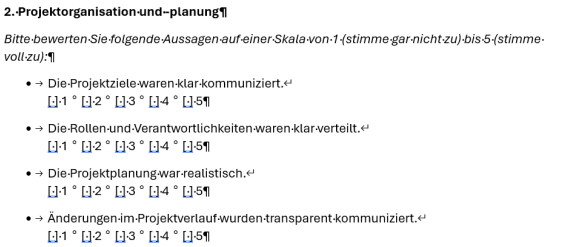


Abb. 1: Einführungsschritte in einem SAP Projekt [1]

Evaluationsmöglichkeiten

Um am Ende eines Projektes die möglichen Fehler, die im Laufe dessen angefallen sind, ordentlich zu dokumentieren bietet sich eine Evaluation an. Hierfür gibt es verschiedene Möglichkeiten zum Beispiel, eine Mitarbeiterbefragung. [4] Hier erfährt man anonym von den Teilnehmern des Projektes, was in ihren Augen positiv und negativ gelaufen ist. Es ist aber auch möglich, eine Fremdfirma zu beauftragen, die sich auf Evaluationen spezialisiert hat, um eine solche Evaluation durchzuführen. Diese Firmen bewerten dann das Projekt an verschiedenen Parametern und geben ein detailliertes Ergebnis zurück.


2.·Projektorganisation·und·-planung¶

Bitte·bewerten·Sie·folgende·Aussagen·auf·einer·Skala·von·1·(stimme·gar·nicht·zu)·bis·5·(stimme·voll·zu):¶

- → Die·Projektziele·waren·klar·kommuniziert.↵
  [·]·1 ° [·]·2 ° [·]·3 ° [·]·4 ° [·]·5¶
- → Die·Rollen·und·Verantwortlichkeiten·waren·klar·verteilt.↵
  [·]·1 ° [·]·2 ° [·]·3 ° [·]·4 ° [·]·5¶
- → Die·Projektplanung·war·realistisch.↵
  [·]·1 ° [·]·2 ° [·]·3 ° [·]·4 ° [·]·5¶
- → Änderungen·im·Projektverlauf·wurden·transparent·kommuniziert.↵
  [·]·1 ° [·]·2 ° [·]·3 ° [·]·4 ° [·]·5¶

Abb. 2: Auszug aus dem Mitarbeiterbefragungsbogen [1]

Ausblick

Da die Arbeit an diesem Zeitpunkt noch nicht vollständig abgeschlossen ist, wird hierfür ein Ausblick gegeben. Im Projekt stehen im nächsten Schritt die Testphase an. Hierfür wurden bereits Tests programmiert, aber diese konnten noch nicht im System implementiert und getestet werden. Im Moment fehlt den Teilnehmern des Projektes leider der Zugriff auf das System, das sich zurzeit in der Wartung befindet. Das wirkt sich ebenfalls negativ auf den Zeitplan aus, der sich dadurch leicht verzögern wird.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] k unbekannt. Der Ablauf eines SAP-Projekts – Damit müssen Sie rechnen. https://www.wsw.de/wsw-wissen/der-ablauf-eines-sap-projekt-damit-muessen-sie-rechnen/, 2025.

[3] k unbekannt. Einführungs- & Projektmethodik. https://www.fis-gmbh.de/de/produkte-leistungen/sap-einfuehrung-und-beratung/sap-implementierung/, 2025.

[4] k unbekannt. Evaluation der Projektarbeit. *Bürger Stiftung Hamburg*, 2025.

[5] k unbekannt. Finanz-Sanktionsliste. https://justiz.de/onlinedienste/finanz_sanktionsliste/index.php, 2025.

[6] k unbekannt. Zollabwicklung. https://simpleclub.com/lessons/kaufmann-frau-fur-spedition-logistikdienstleistung-zollabwicklung, 2025.

Real-Time Suggestions for Optimizing Fleet Distribution of Sharing Services

Ruben Roehner **Mirko Sonntag**

Department of Computer Science and Engineering, Esslingen University

Work carried out at Department of Computer Science and Engineering, Esslingen

Introduction

Shared micromobility services, such as electric scooter sharing, are a large, growing part of urban and suburban mobility. Shared micromobility offers an ecologically friendly and convenient way to bridge short distances. They address the common problem of the "first and last mile gap", which describes the distance between public transport stations and the trip's origin and destination. The growing popularity of micromobility sharing services comes from their flexibility, ease of use, low environmental impact, and the possibility of reducing private car rides. However, service operators face challenges in maintaining high service quality to realize these benefits effectively. Rebalancing, the strategy of relocating vehicles from areas of low demand to areas of high demand, is a critical operational task for maintaining a balanced fleet and thereby ensuring vehicle availability and user satisfaction [5]. Rebalancing strategies are generally categorized into two forms: operator-based rebalancing, where service workers collect and relocate vehicles, and user-based rebalancing, where users are incentivized to perform rebalancing tasks [4].
This master's thesis aims to develop and implement a hierarchical reinforcement learning (HRL) framework for e-scooter rebalancing, evaluate its performance against existing methods, and provide a system capable of offering real-time suggestions for fleet optimization through rebalancing operations.

Background and Related Works

The data underlying this thesis was collected from a General Bike Sharing Specification (GBFS) feed that publishes the live locations of e-scooters from sharing services operating in the greater Stuttgart area [2]. These services utilize a free-floating approach, permitting users to pick-up and drop-off vehicles at any location within the operating area, without restriction to specific parking stations. The raw GBFS data undergoes a pre-processing stage to extract pick-up and drop-off events. These events were subsequently spatially and temporally aggregated using the H3 grid system [1] and hourly time bins for analysis, as well as for training and evaluating the proposed rebalancing framework.
Analysis of the collected GBFS data revealed distinct spatio-temporal patterns in vehicle movement. For instance, a significant net imbalance was observed during morning hours (3-8 a.m.), characterized by a flow of e-scooters from suburban areas towards Stuttgart's city center. Figure 1 illustrates this pattern, with H3 grid zones color-coded to indicate undersupply (red) or oversupply (blue) of vehicles. Conversely, the evening hours demonstrate an opposite trend, with most trips originating in the city center and ending in suburban areas. This observed imbalance underscores the importance of rebalancing.

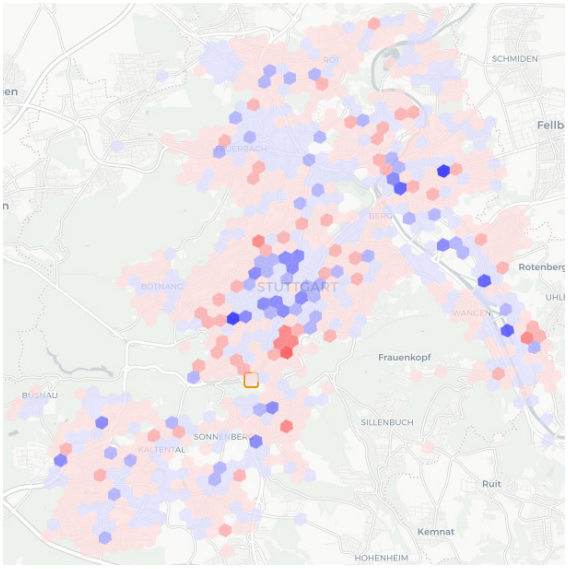


Fig. 1: E-scooter imbalance hot spots from 3 a.m. to 8 a.m. Red coloring indicates undersupply, and blue coloring indicates an oversupply of vehicles. [8]

Previous research on rebalancing algorithms has mainly focused on user- or operator-based strategies in isolation, often lacking a comprehensive framework that integrates both. Common approaches to rebalancing include heuristic methods, mathematical optimization, and reinforcement learning (RL). Several studies indicate that RL approaches can outperform traditional heuristics and mathematical optimization techniques, such as Mixed Integer Linear Programming (MILP) or simple greedy heuristics [7].

Proposed Framework

A hierarchical reinforcement learning (HRL) framework is proposed to address the e-scooter rebalancing challenge. The HRL framework comprises two tiers. The first tier, the Regional Distribution Coordinator, manages inter-community imbalances by determining the target number of vehicles for each defined community. These imbalances are addressed through operator-based rebalancing, where service workers collect vehicles from communities with an oversupply and redistribute them to communities experiencing an undersupply. The second tier consists of multiple User-Incentive Coordinators, each assigned to a specific community, to manage intra-community imbalances (i.e., imbalances within a single community). These coordinators aim to promote an equal distribution of vehicles by offering incentives to users for returning vehicles to designated undersupplied zones at the end of their trips.

A dedicated demand forecasting module supports both the Regional Distribution Coordinator and the User-Incentive Coordinators. This module provides demand predictions as part of the state for both tiers, enhancing the effectiveness of rebalancing operations by allowing the system to anticipate future demand. This modular design ensures that both tiers optimize rebalancing decisions based on consistent future outlooks and separates the demand forecasting task from the RL agents. Additionally, the current vehicle distribution across all zones is provided as input to both agents. Figure 2 illustrates the architecture of this two-tier HRL framework.

A multi-head Deep Q-Network (DQN) [6] is employed for the Regional Distribution Coordinator. Each head corresponds to a community, determining the net target number of e-scooters to be collected from or delivered to that community. The multi-head DQN utilizes Experience Replay to optimize data usage and facilitate the learning of an effective rebalancing strategy. A Proximal Policy Optimization (PPO) [9] algorithm is implemented for the User-Incentive Coordinators. This algorithm allocates incentives to specific zones within a community to encourage users to drop-off vehicles in these targeted zones rather than their originally intended destinations. The primary objective for both RL agents is to maximize the number of satisfied user trips while minimizing the cost associated with rebalancing operations.

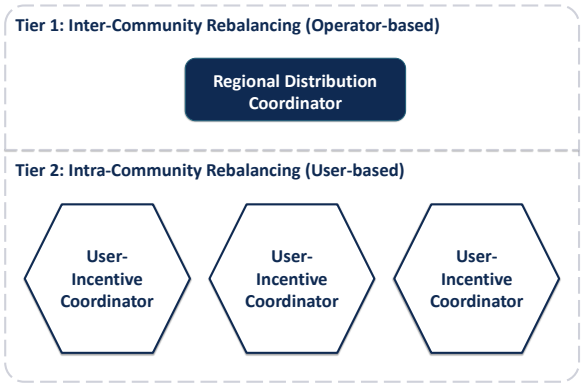


Fig. 2: Conceptual illustration of the two-tier Hierarchical Reinforcement Learning framework for e-scooter rebalancing. [8]

A simulation environment is developed to evaluate the proposed HRL framework. Performance is measured using metrics such as satisfied user demand and the Gini coefficient (to assess the fairness of vehicle distribution) [3]. The results from evaluating the HRL framework are benchmarked against several baselines, including user-based rebalancing only, operator-based rebalancing only, and existing integrated rebalancing frameworks, to demonstrate the effectiveness of the proposed combined hierarchical approach.

Conclusion and Future Work

This thesis introduces a novel hierarchical reinforcement learning framework for addressing the rebalancing problem in shared e-scooter services. The proposed framework integrates operator-based and user-based rebalancing strategies: operator-based rebalancing addresses broader, inter-community vehicle distribution, while user-based rebalancing targets local, intra-community imbalances. The real-time application of this framework has the potential to significantly improve service quality by enhancing vehicle availability and dynamically responding to shifts in user demand. Future enhancements to this HRL framework involve exploring the implementation of alternative RL algorithms for both the Regional Distribution Coordinator and the User-Incentive Coordinators. Promising candidates include Actor-Critic algorithms such as Deep Deterministic Policy Gradient (DDPG) or Advantage Actor-Critic (A2C). Further research could also focus on developing a high-fidelity simulation environment, based on platforms like SUMO (Simulation of Urban Mobility), to enable more realistic evaluation under conditions that closely approximate real-world scenarios.

References and figures

[1] Isaac Brodsky. H3: Uber's Hexagonal Hierarchical Spatial Index. https://www.uber.com/en-DE/blog/h3/, 2018.

[2] MobiData BW. Gebündelte Daten E-Scooter-Sharing Baden-Württemberg - MobiData BW® (GBFS-Feed). https://www.mobidata-bw.de/dataset/escootch, 2025.

[3] Matteo Cederle et al. A Fairness-Oriented Reinforcement Learning Approach for the Operation and Control of Shared Micromobility Services. https://arxiv.org/abs/2403.15780, 2024.

[4] Elnaz Emami and Mohsen Ramezani. Integrated operator and user-based rebalancing and recharging in dockless shared e-micromobility systems. *Communications in Transportation Research*, 4:100155, 2024.

[5] Sujae Kim et al. Optimal Rebalancing Strategy for Shared e-Scooter Using Genetic Algorithm. *Journal of Advanced Transportation*, 1:2696651, 2023.

[6] Volodymyr Mnih et al. Human-level control through deep reinforcement learning. *nature*, 518:529–533, 2015.

[7] Ling Pan et al. A Deep Reinforcement Learning Framework for Rebalancing Dockless Bike Sharing Systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:1393–1400, 2019.

[8] Own representation.

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

Entwicklung einer webbasierten E-Mail-Verwaltungssoftware zur Optimierung der Nutzungserfahrung und Effizienz.

Phileas Roth **Astrid Beck**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Steuerbot GmbH, Welfenstraße 19, 70736 Fellbach

Einleitung

Das Unternehmen Steuerbot hat eine App entwickelt, die Nutzern hilft, ihre Steuererklärung auf einfache Weise zu erstellen. Mit Hilfe einer Chat-Schnittstelle führt die App durch alle relevanten Themen und gibt klare Erklärungen und gezielte Hilfestellungen. Nach dem Ausfüllen der Steuererklärung erhält der Nutzer automatisch eine E-Mail mit Informationen wie dem aktuellen Steuerbescheid. Diese E-Mails bestehen aus einfachen HTML-Dateien, die derzeit vom Steuerbot-Team manuell programmiert werden. Für die Erstellung neuer E-Mails gibt es ein eigenes Git-Projekt, für dessen Nutzung technische Vorkenntnisse erforderlich sind. Ziel dieser Arbeit ist es, ein Tool zu entwickeln, das diesen Prozess vereinfacht, technische Kenntnisse überflüssig macht und die Erstellung und Verwaltung von E-Mails effizienter gestaltet.

Zielsetzung

Ziel dieses Projekts ist es, eine webbasierte Software zu entwerfen und zu entwickeln, die die Arbeit mit Steuerbot-E-Mails deutlich vereinfacht. Dies soll die Nutzererfahrung verbessern und die Effizienz bei der Bearbeitung von E-Mails erhöhen. Die Software basiert auf dem bestehenden E-Mail-Projekt, so dass die bestehende Struktur und die bereits entwickelten Komponenten für die E-Mail-Erstellung weiter genutzt werden können. Der wesentliche Unterschied besteht darin, dass die Teammitglieder die E-Mails nicht mehr manuell programmieren müssen, sondern sie direkt über eine Weboberfläche mit Drag-and-Drop-Komponenten gestalten und anpassen können.

Theoretische Grundlagen

MJML ist eine deklarative Markup-Sprache, die speziell zur Vereinfachung der Erstellung von responsiven E-Mails entwickelt wurde. Mithilfe einer semantischen Syntax und einer umfangreichen Standardbibliothek abstrahiert MJML die Komplexität der herkömmlichen HTML-E-Mail-Entwicklung und erzeugt standardkonformes, responsives HTML. Als Open-Source-Lösung basiert MJML auf aktuellen Best Practices und ermöglicht eine effiziente, wartungsarme Umsetzung moderner E-Mail-Layouts [1]. MJML-react ist eine Open-Source-Bibliothek, die es ermöglicht, responsive E-Mails in React mit einer deklarativen Komponentenstruktur zu erstellen. Sie abstrahiert die Komplexität der MJML-Syntax in typisierte React-Komponenten und ermöglicht so eine nahtlose Integration in moderne Webentwicklungs-Workflows. Die automatische Generierung aktueller MJML-Komponenten und zusätzliche Funktionen wie Typsicherheit und HTML-Post-Processing sorgen für eine effiziente, wartbare und kompatible E-Mail-Entwicklung [3].

Funktionsweise der Vorschau und Übersetzung

Um eine E-Mail in der Vorschau anzuzeigen, werden mehrere Schritte durchgeführt (siehe Abb. 1). Da die E-Mails nicht nur lokal verarbeitet, sondern auch gespeichert werden sollen, werden die zugehörigen Daten in einer MongoDB-Datenbank abgelegt. Wird eine E-Mail geöffnet, wird sie zunächst aus der Datenbank geladen. Jede E-Mail wird dort als JSON-Objekt gespeichert, dessen zentraler Bestandteil ein Baum ist, der die Struktur der E-Mail beschreibt. In diesem Baum werden die einzelnen Komponenten und ihre jeweiligen Eigenschaften definiert, die später die Grundlage für die HTML-E-Mail bilden. Nach dem Laden wird die E-Mail im Frontend als interaktive Baumstruktur angezeigt. Der Benutzer kann die Eigenschaften einer Komponente bearbeiten, indem er auf sie klickt. Währenddessen läuft im Hintergrund ein Übersetzungsprozess ab: Die Baumstruktur wird rekursiv in MJML-React-Code umgewandelt. Dazu

wird jede Komponente ausgehend von den untersten Kindobjekten mit Hilfe einer speziell entwickelten Konfigurationsdatei bearbeitet. Diese enthält nicht nur die Zuordnung zu MJML-React-Komponenten, sondern auch zusätzliche Informationen wie Namen und Beschreibungen, die dem Benutzer helfen, die Bedeutung der jeweiligen Komponente besser zu verstehen. Sobald der MJML-React-Code generiert wurde, wird er in reinen MJML-Code umgewandelt. Dieser wird dann in HTML übersetzt, damit die E-Mail bei jeder Änderung der Eigenschaften live als Vorschau angezeigt werden kann. Die fertige HTML-E-Mail kann exportiert werden. Dieses Format wird benötigt, um die E-Mails korrekt in die Steuerbot-App zu integrieren und an die Nutzer zu versenden.

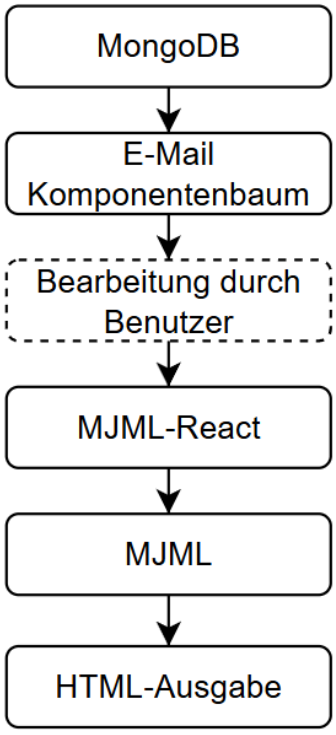


Abb. 1: Übersetzungspipeline einer E-Mail [2]

Aufbau der Anwendung

Die Anwendung besteht aus zwei separaten Ansichten. Nach dem Aufruf der Webanwendung wird zunächst eine Übersichtsseite angezeigt, die die vorhandene Ordner- und E-Mail-Struktur darstellt (siehe Abb. 2). In dieser Ansicht können die Ordner und E-Mails in Listenform verwaltet werden. Der Inhalt der E-Mails wird in verkürzter Form dargestellt, da nicht alle Detailinformationen für die Übersichtsdarstellung benötigt werden. Es ist möglich, neue Einträge hinzuzufügen, bestehende zu bearbeiten oder zu löschen. Eine integrierte Suchfunktion erleichtert zudem die Navigation in umfangreichen Datenstrukturen.

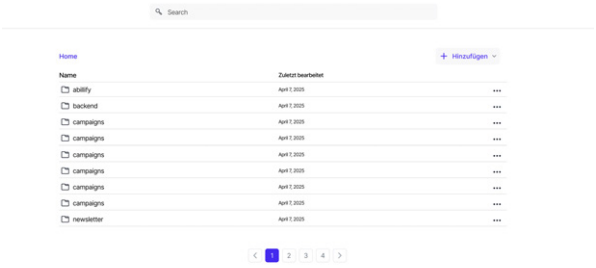


Abb. 2: Designentwurf der Übersichtsseite [2]

Durch Auswahl einer E-Mail aus der Liste wird die Detailansicht der Anwendung aufgerufen, in der der vollständige Inhalt der jeweiligen E-Mail angezeigt wird (siehe Abb. 3). Am oberen Rand dieser Seite befinden sich Bedienelemente wie die Anzeige des E-Mail-Namens, Schaltflächen zum Umschalten zwischen verschiedenen Sprachversionen und zwischen Desktop- und Mobilansicht sowie eine Export-Schaltfläche, mit der die E-Mail als HTML-Datei in Deutsch und Englisch heruntergeladen werden kann. Der zentrale Bearbeitungsbereich ist in zwei Spalten unterteilt, auf der linken Seite befindet sich die hierarchische Baumstruktur der E-Mail-Komponenten. Wird eine Komponente ausgewählt, öffnet sich ein Formular zur Bearbeitung der zugehörigen Eigenschaften. Auf der rechten Seite wird die E-Mail gleichzeitig als Vorschau gerendert, so dass Änderungen sofort sichtbar sind.

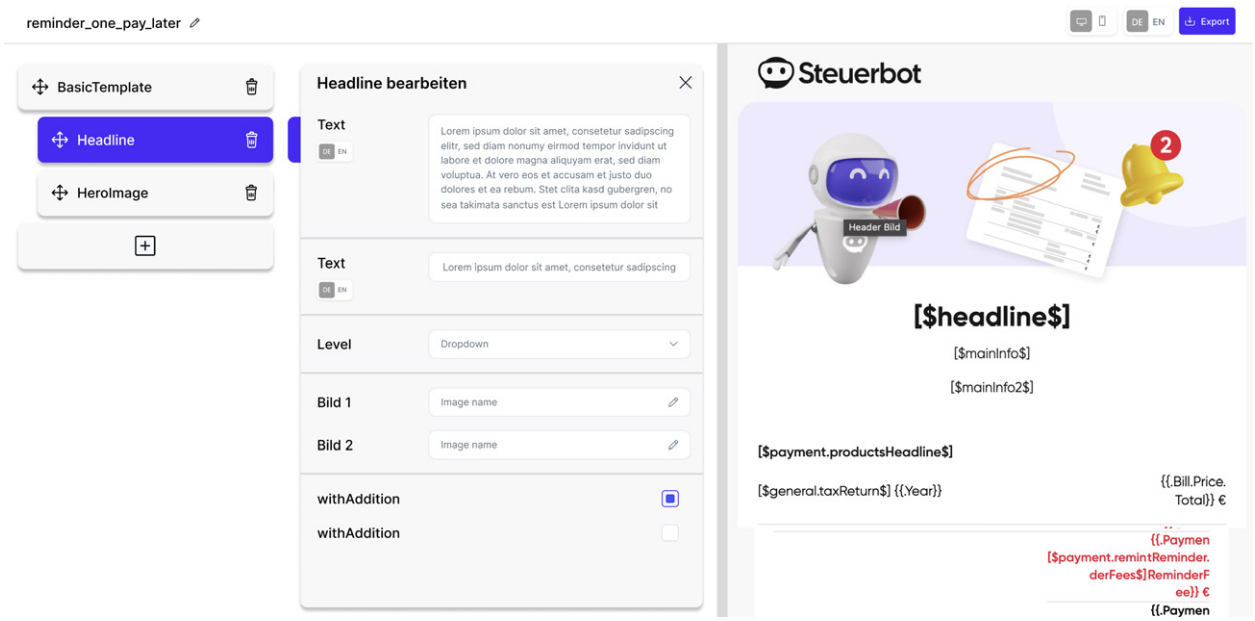


Abb. 3: Designentwurf der Vorschauseite [2]

Ausblick

Da die Liste der Ordner und gespeicherten E-Mails potenziell sehr groß werden kann, ist die Integration einer Paginierungslogik geplant (siehe Abb. 2). Diese soll sicherstellen, dass nur ein Teil der Daten aus der Datenbank geladen wird, um die Datenmenge zu reduzieren und die Leistung der Anwendung zu verbessern. Sobald die Anwendung einen funktionsfähigen Zustand erreicht hat, sind Benutzertests mit der Zielgruppe geplant, zum Beispiel mit dem Marketingteam, das regelmäßig Newsletter-E-Mails erstellt. Die Ergebnisse dieser Tests werden in die weitere Entwicklung einfließen, insbesondere in die Erstellung von benutzerfreundlichen Anleitungen. Diese soll zeigen, wie einzelne Komponenten genutzt werden können und wo ihr Einsatz sinnvoll ist. Auch die Unterstützung von Hell- und Dunkelmodus ist geplant. In Zukunft soll es möglich sein, direkt in der Vorschau zwischen den beiden Darstellungsmodi in Echtzeit umzuschalten.

Literatur und Abbildungen

[1] Maxime Brazeilles. MJML – The only framework that makes responsive email easy. https://mjml.io/, 2024.

[2] Eigene Darstellung.

[3] Michal Jez. MJML-React – React Components for MJML. https://github.com/Faire/mjml-react, 2025.

Evaluierung der Systemgrenzen im Sichtfeld der Sensorik eines Notbremsassistenten für Nutzfahrzeuge

Nick Saurer **Thao Dang**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Daimler Truck AG, Untertürkheim

Einleitung

Unfälle mit Lastkraftwagen (LKW) haben aufgrund ihrer enormen Masse und hohen kinetischen Energie oft schwerwiegende Folgen. [4] Im Jahr 2023 waren LKW in Deutschland an 23.929 Unfällen mit Personenschaden beteiligt, was die Risiken im Straßenverkehr verdeutlicht. [6] Die zunehmende Verbreitung von Fahrerassistenzsystemen spielt eine entscheidende Rolle bei der Unfallvermeidung. Eine Studie der BASt ergab einen Rückgang tödlicher Unfälle mit schweren Nutzfahrzeugen um etwa 30 Prozent aufgrund von Notbremsassistenten im Zeitraum von 2002 bis 2015. [4] Besonders durch die EU-Verordnung Nr. 2019/2144, die seit Juli 2022 für schwere Nutzfahrzeuge verpflichtend die Weiterentwicklung von Notbrems- und Abbiegeassistenzsystemen vorschreibt. [1] Angesichts dieser Risiken und der regulatorischen Entwicklungen ist eine vertiefte Auseinandersetzung mit den Leistungsgrenzen dieser Systeme bedingt durch Sensorik und Algorithmen unerlässlich, um ihre Sicherheit zu bewerten und Verbesserungspotenziale aufzuzeigen.

Ziel der Arbeit

Ziel dieser Arbeit ist die Untersuchung eines Notbremsassistenten hinsichtlich Fehlverhaltens, das durch ungenauer oder falscher Sensordaten verursacht wird. Im Fokus dieser Analyse der Systemgrenzen steht die Identifizierung von Szenarien, in denen unzureichende oder fehlerhafte Sensordaten zu einer inkorrekten Objektbewertung und somit zu kritischen Fehlreaktionen, wie ausbleibenden notwendigen Bremsungen des Systems führen. Des Weiteren sollen durch die detaillierte Betrachtung der sensorischen Informationsverarbeitung Bereiche mit Leistungseinbußen identifiziert und die Grenzen der aktuellen Sensorik bestimmt werden, um Ansatzpunkte für Optimierungen aufzuzeigen.

Erprobungsfahrten

Die Validierung von Notbremsassistenten erfolgt durch Erprobungsfahrten unter realistischen Bedingungen. Dabei wird das zu testendes Fahrzeug gezielt auf ein standardisiertes Global Vehicle Target (GVT) zugeführt. Das GVT ist ein Soft-Target, das die Eigenschaften eines typischen Pkw gegenüber optischen Sensoren sowie Radar- und Lidarsystemen nachbildet. Dank seines weichen Materials können Kollisionen ohne Schäden am Testfahrzeug simuliert werden. Ein solches Soft-Target ist in Abbildung 1 dargestellt.

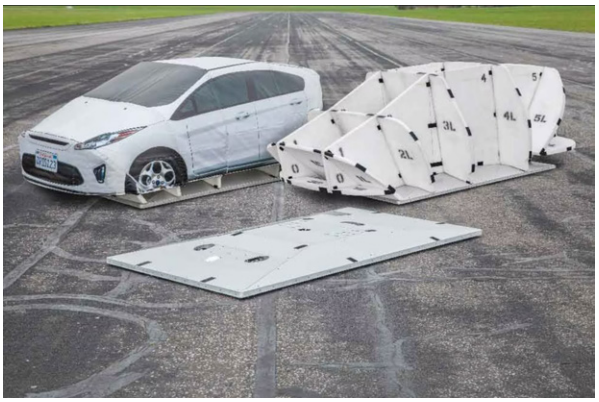


Abb. 1: Soft-Target (GVT) mit eingebautem DGPS [5]

Zur genauen Messung der Fahrzeugbewegung und Position wird Differenzielles GPS (DGPS) eingesetzt. DGPS verbessert die Positionsgenauigkeit durch Korrekturdaten, indem es Pseudostreckenfehler eliminiert. Dies ermöglicht eine präzise Analyse der Fahrzeugposition und -bewegung während der Testfahrten. [2] Durch die Aufzeichnung der Erprobungsfahrten werden sämtliche Sensordaten und Ausgangsgrößen des Systems dokumentiert. Zudem stehen die DGPS-Daten des Global Vehicle Target (GVT) zur Verfügung, wodurch eine präzise Analyse der Systemreaktion hinsichtlich der Sensorik möglich wird.

Vector CANape

Die in den Testfahrten gesammelten Messdaten können mit Hilfe des Softwaretools Vector CANape detailliert ausgewertet werden. Diese Software ermöglicht eine präzise Analyse der Sensordaten und Systemausgangsgrößen, wobei spezielle Erweiterungen für ADAS-Funktionen verfügbar sind. In einer grafischen Oberfläche werden alle erfassten Objekte der Sensorik sichtbar gemacht. Zudem können DGPS-Daten des Global Vehicle Target (GVT) angezeigt werden, wodurch sich Abstände und Positionen einschätzen und sehen lassen. Ergänzend hierzu bietet ein Video-Fenster die Möglichkeit, die Erprobungsfahrten erneut abzuspielen und detailliert mit den Messdaten abzugleichen. Eine beispielhafte grafische Oberfläche und ein Video-Fenster sind in Abbildung 2 zu sehen.

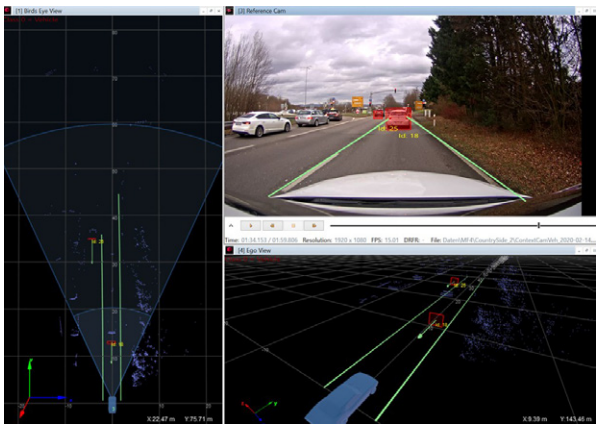


Abb. 2: Beispielhaftes Szenen- und Video-Fenster in CANape [3]

Alle aufgezeichneten Messsignale lassen sich in verschiedenen Analyse-Fenstern zeitgenau im Millisekunden-Bereich betrachten, vergleichen und auswerten. Für weitergehende Untersuchungen stellt CANape in seiner Data-Mining-Umgebung eine Skriptfunktion bereit, mit der Messsignale eingelesen, Bedingungen geprüft sowie Werte umgerechnet oder miteinander verrechnet werden können. [3] Diese umfangreichen Analyseoptionen bilden die Grundlage meiner Bachelorarbeit, um die Systemgrenzen des Notbremsassistenten aus Sicht der Sensorik zu evaluieren.

Konzept

Die Analyse der Systemgrenzen aus Sicht der Sensorik erfolgt innerhalb der Data-Mining-Umgebung, in der das Skript Sensordaten ausliest und die Reaktionsbedingungen des Systems überprüft. Dabei werden aufgezeichnete Messdaten automatisiert verarbeitet und über die Zeit iteriert, sodass eine präzise Auswertung und Verrechnung der einzelnen Signale möglich wird. Ein zentraler Aspekt dieser Analyse ist die Einschätzung der Kritikalität einer Fahrsituation während der Erprobungsfahrt. Hierbei spielen die DGPS-Daten des Global Vehicle Target (GVT) eine entscheidende Rolle. Durch Berechnungen der Kinetik, darunter Geschwindigkeit, Abstände und Beschleunigungen, wird die Situation bewertet und unzureichende Systemreaktionen identifiziert. Nach dieser Einschätzung erfolgt die automatische Auswahl des relevanten Sensorik-Objekts, dessen Sensordaten weitergehend analysiert werden. In den Bereichen der unzureichenden Systemreaktion werden gezielt die Sensorzustände ausgewertet, wobei fehlerhafte Sensorbedingungen dokumentiert und gespeichert werden. Abschließend werden diese fehlerhaften Bedingungen in eine CSV-Datei exportiert, die alle fehlgeschlagenen Bedingungen der Sensorik auflistet. Dadurch lässt sich nachvollziehen, warum das System in bestimmten Situationen nicht reagiert hat oder ob in allen kritischen Bereichen sämtliche Bedingungen erfüllt waren und die Reaktion des Notbremsassistenten als ausreichend bewertet werden kann. Diese detaillierte Analyse bildet die Grundlage zur weiteren Optimierung des Notbremsassistenten und zeigt Systemgrenzen aus Sicht der Sensorik auf.

Ausblick

Die detaillierte Analyse der Testfahrten mit den Sensordaten ermöglicht wertvolle Optimierungen und neue Ansätze zur Verbesserung der Funktionsreaktion des Notbremsassistenten. Durch gezielte Anpassungen können Systeme weiterentwickelt werden, um ihre Reaktionsfähigkeit in kritischen Situationen zu maximieren. Gleichzeitig bietet die Auswertung die Möglichkeit, neu entwickelte Optimierungen zu validieren, indem sie mit bisherigen Systemreaktionen verglichen werden. Dadurch lässt sich bestimmen, ob die Verbesserungen tatsächlich zu einer ausreichenden und zuverlässigeren Funktionalität führen. Ein weiterer bedeutender Aspekt ist die Identifizierung fester Systemgrenzen, an denen die Sensorik an ihre physikalischen und technischen Limits stößt. Diese Erkenntnisse sind entscheidend für zukünftige Entwicklungen, da sie klar definieren, unter welchen Bedingungen die Assistenzsysteme noch einwandfrei arbeiten und ab wann alternative Lösungsansätze erforderlich sind.

Literatur und Abbildungen

[1] EU Europäisches Parlament. Verordnung (EU) 2019/2144 des Europäischen Parlaments und des Rates. https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=celex%3A32019R2144, 11 2019.

[2] Racelogic GmbH. Wie funktioniert DGPS (Differentielles GPS)? https://de.racelogic.support/01VBOX_Automotive/01Allgemeine_Informationen/Knowledge_Base/Wie_funktioniert_DGPS_(Differentielles_GPS)%3F, 07 2021.

[3] Vector Informatik GmbH. CANape Option Driver Assistance. https://www.vector.com/de/de/produkte/produkte-a-z/software/canape/canape-option-driver-assistance/#, 2025.

[4] Rainer Müller-Finkeldei. How high-tech equipment in trucks is saving lives. https://medium.com/transportation-matters/how-high-tech-equipment-in-trucks-is-saving-lives-996473634ef1, 12 2019.

[5] Grant Maloy Smith. Wie werden ADAS-Systeme und autonome Fahrzeuge getestet? https://dewesoft.com/de/blog/testen-von-adas-systemen-und-autonomen-fahrzeugen, 02 2023.

[6] DESTATIS Statistisches Bundesamt. Unfälle von Güterkraftfahrzeugen im Straßenverkehr 2020. https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Verkehrsunfaelle/Publikationen/Downloads-Verkehrsunfaelle/unfaelle-gueterkraftfahrzeuge-5462410207004.pdf, 01 2022.

Bewertung der Navigationsperformace von Mährobotern durch die Schätzung von Geschwindigkeit und Beschleunigung anhand einer Pose-Zeit-Sequenz

Nuro Savelsberg **Jürgen Koch**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma STIHL AG, Waiblingen

Einleitung

Für autonome Systeme wie unter anderem Mähroboter ist die Navigation ein ausschlaggebendes Qualitätsmerkmal. Gleichzeitig ist die Bewertung der Navigationsperformance mit einigen Schwierigkeiten verbunden. Vor allem bei der Bewertung der Navigationsperformance von Produkten, bei denen die Implementierung der Navigationsstrategie unbekannt ist, ist eine objektive Bewertung bisher kaum möglich. Im Rahmen dieser Arbeit sollen Messwerte eines motorisierten Tachymeters verwendet werden, um eine automatisierte Bewertung der Navigationsperformance von Mährobotern zu ermöglichen.

Motivation

Bisher erfolgt die Bewertung von Mährobotern vor allem anhand der Spezifikationen und anhand des Hardware-Aufbaus. Die Bewertung der Navigationsperformance erfolgt durch eine subjektive Einschätzung basierend auf Beobachtungen und eventuellen Fehlern, wie beispielsweise dem Verlassen des Mähgebiets oder dem unzureichenden Erkennen von Hindernissen. Nur eine objektive und systematische Bewertung würde einen aufschlussreichen Vergleich verschiedener Navigationssysteme ermöglichen und die Betrachtung der Navigationsperformance über mehrere Jahre zulassen. Für solch eine objektive Bewertung existiert bisher kein System.

Aufbau

Die automatisierte Bewertung wird durch das Leica iCON iCS50 Tachymeter ermöglicht. Tachymeter messen die dreidimensionale Position eines Objekts durch eine elektrooptische Distanzmessung und die Erfassung des Horizontal- sowie Vertikalwinkels [2]. Das iCS50 zeichnet sich durch die Fähigkeit aus, zusätzlich auch die Orientierung des Objekts erfassen zu können. Dafür wird eine gepunktete Messkugel verwendet. Anhand der Ausrichtung dieser Punkte zueinander kann die Orientierung der Kugel eindeutig bestimmt werden. Das iCS50 verfügt über motorisierte Achsen, was die automatische Zielverfolgung der Kugel ermöglicht. Dadurch kann die Trajektorie eines Mähroboters während des Mähvorgangs aufgezeichnet werden.

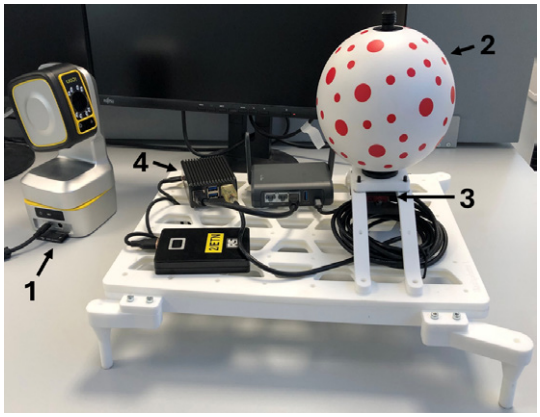


Abb. 1: Versuchsaufbau: 1: Das iCS50, 2: Leica Messkugel, 3: eine Inertial Measurement Unit (IMU) und 4: ein Raspberry Pi. Nur 1 und 2 sind für das System notwendig, während 3 und 4 der Validierung dienen. [3]

Zielsetzung

Um eine objektive Bewertung zu ermöglichen, sollen Kriterien identifiziert werden, die mit einem *score* bewertet werden können. Basierend auf diesen Kriterien soll ein Bericht automatisch erstellt werden, der eine Bewertung der Mähperformance beinhaltet.
Für einige dieser Kriterien sind Werte für die Geschwindigkeit und Beschleunigung in allen sechs Achsen nötig. Schätzungsmethoden, die diese Größen anhand der

vom Tachymeter gemessenen Positions-Zeit-Sequenz schätzen, sollen implementiert und verglichen werden. Um Kriterien implementieren zu können, die das Verhalten während spezifischer Fahrzustände beschreiben, soll die Trajektorie in Abschnitte verschiedener kinematischer Zustände segmentiert und klassifiziert werden. Dabei soll zwischen kinematischen Zuständen, wie Beschleunigung, konstante Geschwindigkeit, Kurvenfahrt und ähnlichen Zuständen, unterschieden werden.
Des Weiteren sollen aus der Analyse der gemähten Fläche weitere Bewertungskriterien hervorgehen. Dabei ist relevant, ob Stellen ausgelassen oder mehrmals befahren wurden. Zusätzlich soll untersucht werden, ob die Räder des Roboters einige Stellen häufig überfahren, da das den Rasen beschädigen könnte. Die Flächenauswertung soll als *plugin* für das Programm QGIS (Quantum Geographic Information System) bereitgestellt werden.

Schätzung von Bewegungsinformationen

Um die Bewegungsinformationen zu schätzen, wurden vor allem verschiedene Versionen des Kalman-Filters (KF) betrachtet. Dabei werden die Werte der einzelnen Achsen unabhängig voneinander geschätzt. Um die Werte zu validieren, wurden Messungen mit einer IMU durchgeführt (Abb. 1). Nachdem die Messwerte des iCS50 in das Koordinatensystem des Roboters umgerechnet wurden, können die Schätzwerte mit den Messwerten der IMU verglichen werden. In Abb. 2 werden die Ergebnisse zweier Schätzungsmethoden verglichen und anhand von Messungen der IMU validiert. Betrachtet werden ein KF zweiter Ordnung und ein Interacting Multiple Mode (IMM) Filter. Letzterer beinhaltet mehrere KF mit verschiedenen Konfigurationen. Das Ergebnis des IMM-Filters setzt sich aus den Ergebnissen der einzelnen Filter zusammen, wobei die Gewichtung davon abhängig ist, wie groß die Unterschiede der einzelnen Ergebnisse zu den Messwerten sind [1]. Beide Methoden erreichen eine hohe Übereinstimmung mit den Validierungsdaten. Der IMM-Filter reagiert schneller auf Sprünge in der Eingangsgröße und ist gleichzeitig robuster gegenüber Messrauschen, das Ergebnis des IMM-Filters ist dennoch nur geringfügig besser als das des KF zweiter Ordnung. Insgesamt könnte der KF zweiter Ordnung dennoch bevorzugt werden, da der IMM-Filter deutlich mehr Rechenaufwand benötigt und bei sehr großem Messrauschen instabil werden kann.

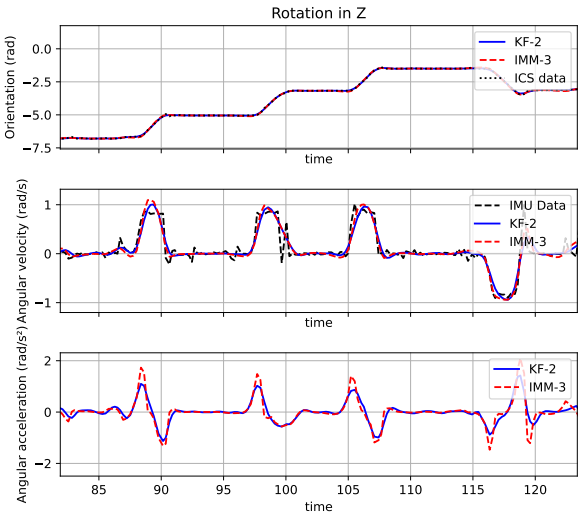


Abb. 2: Vergleich zweier Schätzungsmethoden. Oben Rotation um die z-Achse, mittig Winkelgeschwindigkeit inklusive IMU-Messungen (schwarz) und unten Winkelbeschleunigung. In Rot das Ergebnis eines IMM-Filters, in Blau Ergebnis eines KF 2. Ordnung. [3]

Trajektorie-Klassifizierung

Anhand der geschätzten Geschwindigkeits- und Beschleunigungsdaten können mithilfe von Schwellwerten Abschnitte verschiedener kinematischer Zustände identifiziert und klassifiziert werden. Betrachtet wird die Fahrt- und Rotationsgeschwindigkeit des Roboters, sowie dessen Beschleunigung. Die Klassifizierung in fünf Zustände geht aus einer logischen Verknüpfung der drei betrachteten Größen hervor. In Abb. 3 ist die Klassifizierung eines Mähvorgangs zu sehen. Zunächst fährt der Roboter in geraden Bahnen, bis er auf ein Hindernis stößt, anschließend folgt er dem Begrenzungsdraht. Es wird zwischen einem *zero-turn* und Kurvenfahrt unterschieden. Während eines *zero-turn* muss die Fahrtgeschwindigkeit unter einem Schwellwert sein, während sie bei der Kurvenfahrt darüber ist.

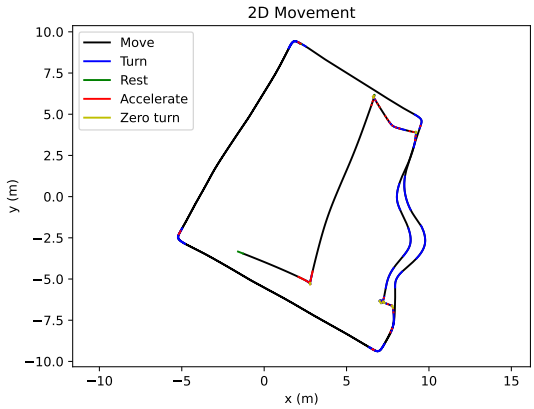


Abb. 3: Trajektorie-Klassifizierung in kinematische Zustände aus der Vogelperspektive über dem Testbereich. Der Pfad des Roboters ist anhand des Zustands farblich gekennzeichnet. [3]

Flächenanalyse

Für die Flächenanalyse soll sowohl die Überlappung des Mähstreifens als auch die der Räder des Roboters bei aufeinanderfolgenden Bahnen analysiert werden. Das ermöglicht einen Vergleich der Abdeckungseffizienz und Überlappung. Außerdem können Risiken der Beschädigung des Rasens durch wiederholten Kontakt mit den Rädern identifiziert werden. Um hierbei die Betrachtung über einen langen Zeitraum zu ermöglichen, soll eine *heatmap* erstellt werden, an der abzulesen ist, wie oft das Mähwerk bzw. die Räder einen Bereich überfahren haben. Durch einen Vergleich der eingelernten Fläche mit der tatsächlich vom Mähwerk überfahrenen Fläche kann eine Aussage über stehen gelassene Flächen getroffen werden. Um die Verwaltung verschiedener Fahrten sowie die Definitionen zu mähender Flächen und Hindernissen zu erleichtern, soll die Flächenauswertung als *plugin* für das Programm QGIS bereitgestellt werden. QGIS ist ein *open-source* und frei zugängliches *geographic information system* (GIS), das zur Verarbeitung und Visualisierun von Karten und räumlichen Daten verwendet wird. Abb. 4 zeigt die Überlappung des Mähwerks bei einer Fahrt auf kleiner Fläche.

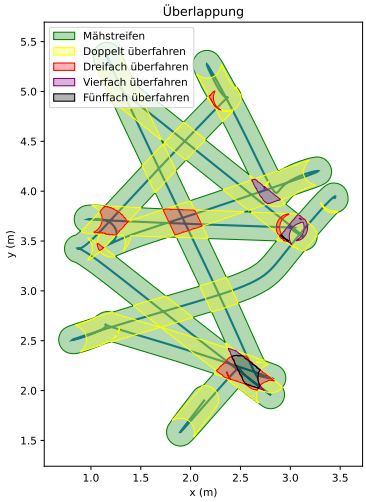


Abb. 4: Analyse der vom Mähwerk abgedeckten Fläche. Farbliche Kennzeichnung, wie oft eine Fläche überfahren wurde. [3]

Ausblick

Während die Implementierung der Zustandsschätzung und der Trajektorie-Klassifizierung abgeschlossen ist, sind weitere Arbeiten im Bereich der Flächenanalyse nötig. Diese Analyse soll mithilfe eines *plugins* für das Programm QGIS visualisiert werden. Des Weiteren muss die automatische Berechnung von *scores* für einige Bewertungskriterien sowie die Erstellung des Bewertungsberichts implementiert werden. Abschließend soll das System in verschiedenen Bedingungen und beim Einsatz unterschiedlicher Mähroboter erprobt werden.

Literatur und Abbildungen

[1] Henk A. P. Blom and Yaakov Bar-Shalom. The interacting multiple model algorithm for systems with Markovian switching coefficients. *IEEE Transactions on Automatic Control*, 33:780–783, 2002.

[2] Ansgar Brunn. Zwangsfreie Richtungsmessungen. *Zeitschrift für Geodäsie, Geoinformation und Landmanagement*, 1:45–51, 2024.

[3] Eigene Darstellung.

Multikriterien-Optimierung von Routen unter Integration von Echtzeit-Ampeldaten und dynamischer Gewichtskalibrierung

Viola Schaefer **Mirko Sonntag**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

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

Einleitung

Moderne Navigationssysteme kalkulieren die optimalsten Routen unter Miteinbeziehung unterschiedlicher Kriterien. Diese Kriterien kann man hauptsächlich als statisch und dynamisch bezeichnen. Statische Kriterien beinhalten Elemente wie Straßentypen und Distanzen, während dynamische Kriterien Echtzeitverkehrsdaten wie Staus und Geschwindigkeitsbegrenzungen betrachten. Ein Element, das bisher nicht in die Berechnungen mit einfließt, sind Ampelschaltungen. Wiederholtes Anhalten und Abfahren wirkt sich negativ auf den Kraftstoffverbrauch, die produzierten Emissionen und den Verkehrsfluss im Allgemeinen aus. Die Einbeziehung von Ampelschaltungen in die Optimierung des Routing-Verfahrens, sowie die Berücksichtigung von Ampelphasen soll damit untersucht und eine passende Implementierung umgesetzt werden.

Grundlagen

Die Routenberechnung unterliegt dem Prinzip, dass unerwünschte Elemente, wie etwa Stauungen, mit einer *Penalty* belegt werden und die optimalste Route dementsprechend die Route mit der geringsten Gesamt-Penalty ist. Eine *Penalty* sind hierbei zusätzliche Kosten, die einer Route zugeschrieben werden. Der Algorithmus, der für das Navigationssystem in dieser Arbeit verwendet wird, ist der A*-Algorithmus, der 1968 initial publiziert worden ist. [1] Er kann als Erweiterung des Dijkstra-Algorithmus angesehen werden, indem er zur gegebenen Kostenfunktion g(x) des Dijkstra-Algorithmus eine heuristische Schätzfunktion h(x) implementiert, die die geschätzten Kosten von einem beliebigen Knoten A zum Zielknoten beinhaltet. An diese Funktion „wird die Anforderung gestellt, dass diese niemals die tatsächlichen Kosten überschätzen darf“. [2] Somit ergibt sich für den A*-Algorithmus die Funktion $f(x) = g(x) + h(x)$, wie in Abbildung 1 dargestellt.

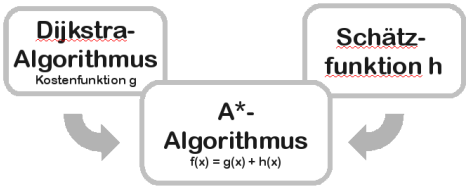


Abb. 1: Abbildung 1: A*Algorithmus [2]

Technologien

Die Technologie, die dafür verwendet werden soll, ist hauptsächlich die Routing-Engine *Valhalla*. *Valhalla* ist eine Open-Source Routing-Engine, die den A*-Algorithmus zur Routenberechnung verwendet. Wie in Abbildung 2 dargestellt ist, setzt sich diese Berechnung aus verschiedenen Faktoren zusammen. Die Karte in *Valhalla* wird aus den *Valhalla-Tiles* zusammengesetzt, die *Edges* beinhalten, also Straßenverläufe. Für diese Arbeit relevant ist der Schritt *Path Finding Using Dynamic Costing (sif)*, da Ampelschaltungen unter die Kategorie der dynamischen Kosten fallen.

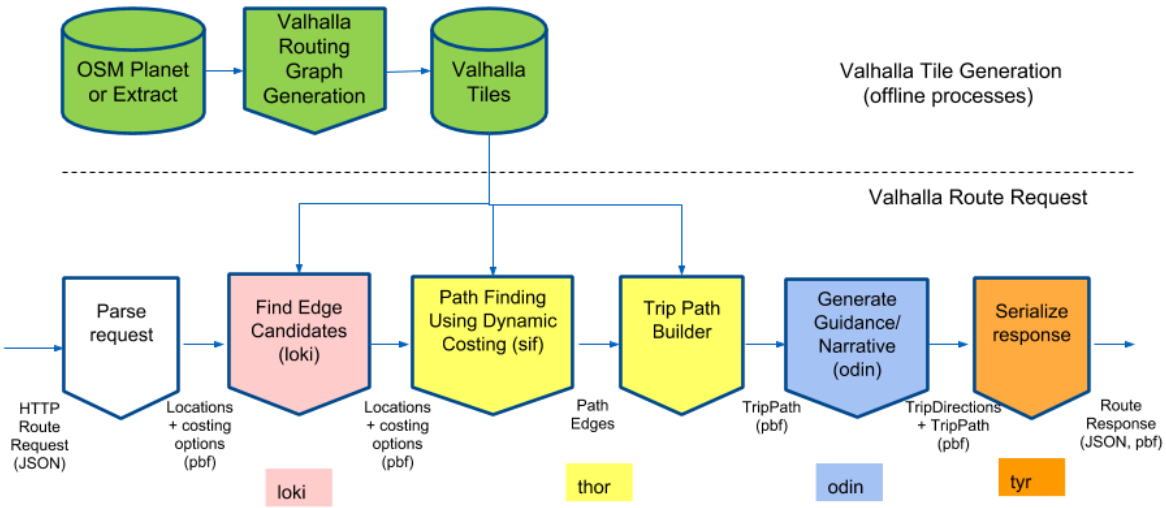


Abb. 2: Abbildung 2: Routenberechnung in Valhalla [3]

Valhalla ist hauptsächlich in der Programmiersprache *C++* geschrieben, womit die Implementierung dem folgen wird. Für die graphische Darstellung von *Heat-Maps*, die noch näher erläutert werden, müssen spezielle *C++-Bibliotheken* verwendet werden.

Ansatz

Bisherige Ansätze sehen es vor, eine Ampelschaltung im Einzelnen zu betrachten und somit jeder Ampelschaltung eine eigene *Penalty* zuzuschreiben. In dieser Arbeit sollen stattdessen *Penalty-Zones* implementiert werden. Anstatt eine einzelne Ampelschaltung zu betrachten, soll das Vorkommen an Ampelschaltungen in einer vordefinierten Zone mithilfe von *Heat-Maps* gemessen werden und die Zone je nach Vorkommen mit einer *Penalty* belegt werden. Dabei wird bei den Ampelschaltungen zwischen reinen Fußgängerampeln und anderen Ampeltypen unterschieden. Reine Fußgängerampeln werden ausschließlich von Fußgängern genutzt und betätigt und die Wahrscheinlichkeit an einer stehen zu bleiben, kann für den Großteil der Zeit vernachlässigt werden. Ampelschaltungen an Kreuzungen hingegen folgen einem bestimmten Ampelphasenzyklus und müssen demnach als potentieller Verzögerungsfaktor berücksichtigt werden. Dementsprechend muss bei der erstellten *Heat-Map* initial eingestellt werden, dass Fußgängerampeln nicht miteinbezogen werden sollen. Würde man nun eine Route berechnen, müsste überprüft werden, ob sie sich in einer *Penalty-Zone* befindet und welche Gewichtung diese Zone aufweist.

Ausblick

In dieser Arbeit ist der Fokus auf der Ausarbeitung einer Implementierung, die durch die Miteinbeziehung von Ampelschaltungen in die Routenberechnung eine erfolgreiche Optimierung des Routings gewährleistet. Dabei müssen andere Kriterien wie Manöverkosten, wie zum Beispiel mehrfaches Abbiegen, und Straßentypen ebenfalls bedacht werden. Würde der Algorithmus eine Route durch ein Wohngebiet führen, anstatt den effizienteren Weg über die Hauptstraße zu wählen, weil dieser zwei Ampelschaltungen beinhaltet, wäre die Optimierung als gescheitert zu betrachten. Während die Erstellung von *Penalty-Zones* den Ampelschaltungen nicht übermäßige Gewichtung zuschreiben würde, müssen andere Faktoren zusätzlich explizit betrachtet werden. So kann zum Beispiel die Tageszeit einen Einfluss auf den Wert einer Route haben, da *Rush-Hours* häufig Stauungen produzieren. Weiterhin können noch Echtzeit-Ampeldaten in die Gewichtung der *Penalty-Zones* miteinfließen. Echtzeit-Ampeldaten sind bisher nur begrenzt vorhanden, weshalb eine zuverlässige Miteinbeziehung bisher noch nicht als ausschließlicher Faktor in der Optimierungsberechnung umsetzbar ist.

Literatur und Abbildungen

[1] Peter Hart, Nils Nilsson, and Bertram Raphael. A Formal Basis for the Heuristic Determination of Minimum Cost Paths. *IEEE Transactions on Systems Science and Cybernetics*, 4:100–107, 1968.

[2] Christoph Hoppe. Softwarepraktikum - A*-Algorithmus. https://pille.iwr.uni-heidelberg.de/~astar01/textbody.html, 10 2008.

[3] David Nesbitt. Overview of how routes are computed - Valhalla Docs. https://valhalla.github.io/valhalla/, 05 2018.

Nutzerzentrierte Entwicklung eines Konzepts für das User Interface des Workspaces innerhalb des Bosch Semantic Stack - Anforderungsanalyse durch qualitative Nutzerinterviews zur Erstellung und Validierung eines Low-Fidelity Prototyps

Lea Jaqueline Scherrbacher **Jörg Nitzsche**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Robert Bosch Manufacturing Solutions GmbH, Stuttgart-Feuerbach

Einleitung

In der Industrie 4.0 entstehen durch die intelligente Vernetzung von Maschinen und Produktionsprozessen große Datenmengen. Dazu zählen beispielsweise Sensordaten von Maschinen, Standortdaten der Transportwege oder Nutzungsdaten wie der Verschleißgrad von Batterien. Diese Daten spielen eine Schlüsselrolle bei der Optimierung von Produkten und Herstellungsprozessen wie auch der Implementierung von Fehlerbehandlungsroutinen. Darüber hinaus entscheiden sie über die Erfüllung gesetzlicher Anforderungen an Nachhaltigkeit und Transparenz der Lieferketten, was die eingesetzten Rohstoffe und Emissionen der Produktion als auch die Entwicklung verbesserter Recyclingmethoden, etwa für Akkus und Batterien, umfasst. [4] Die genannten Daten bleiben jedoch häufig aufgrund organisatorischer Barrieren, inkompatibler Systeme und unterschiedlicher Datenformate in verteilten Silos verborgen, was die weiterführende Nutzung einschränkt. [9], [10] Die Softwarelösungen des *Bosch Semantic Stack* adressieren dieses Problem, indem sie die zentrale Erfassung, Aufbereitung, Speicherung und kontextuelle Einordnung der Daten ermöglichen. Dadurch wird leicht zugängliches Wissen über Produkte und Produktionsabläufe erzeugt und in sogenannten *„Insights-Modulen"* für den Endanwender verständlich aufbereitet sowie über standardisierte APIs in Form von semantisch modellierten Daten zur weiteren Nutzung bereitgestellt. [6], [5]

Problemstellung und Relevanz

Der Semantic Stack steht vor der Herausforderung, große Datenmengen aus verschiedenen Anwendungen über eine grafische Oberfläche effizient und intuitiv nutzbar zu machen. Dabei entscheidet die Optimierung der *Usability und User Experience* (UUX, Gebrauchstauglichkeit und Nutzungserlebnis) über den Erfolg der Softwarelösung, denn dieser schlägt sich in Form höherer *Net Promoter Scores* (Wahrscheinlichkeit der Weiterempfehlung durch Kunden) oder steigender *Conversion Rates* (Umwandlung von Interessenten in Kunden) langfristig in Absatz und Umsatz nieder. [8] Vor diesem Hintergrund werden im Rahmen der Arbeit spezifische UX-Herausforderungen innerhalb des Bosch Semantic Stack untersucht. Dabei gilt es zunächst die vielfältigen Nutzergruppen und deren Bedürfnisse zu identifizieren. Der weitere Fokus liegt anschließend auf drei zentralen Aspekten: der dezentralisierten Struktur der Module und Tools, dem daraus resultierenden zeitaufwändigen Zugang zu diesen sowie dem Fehlen eines zentralen, zielgruppen- und bedarfsgerechten Einstiegs in das Angebot des Semantic Stack.

Aufbau und Funktion des Bosch Semantic Stack

Der Bosch Semantic Stack basiert auf der Technologie der *digitalen Zwillinge*, die physische Produktinstanzen virtuell repräsentieren. In Abbildung 1 ist der vereinfachte, schematische Aufbau des Bosch Semantic Stack dargestellt. Das *Semantic Data Lakehouse* stellt die Basis des Stacks dar und ermöglicht die Integration von Daten aus verschiedenen Quellen. Durch *Ontologien* werden die Daten und Konzepte der Produkte standardisiert beschrieben und deren Beziehungen abgebildet. [1] Innerhalb des *Knowledge Layer* können alle Fakten zu einem Thema in Form von vernetzten *Wissensgraphen* kontextualisiert dargestellt werden.

[7] Dies ermöglicht später die Analyse des gesamten Lebenszyklus eines Produkts und seiner Bauteile mithilfe der Insights-Module. Auch Beziehungen zwischen verschiedenen Produktinstanzen und Produktgruppen können dadurch sichtbar gemacht werden. [6] In der *Digital Twin Registry* werden die IDs der digitalen Zwillinge gespeichert, deren Eigenschaften in Form von Aspekten zusammengefasst werden. Beispiele dafür sind das Gewicht mit der Einheit *Gramm* oder Liegezeiten zwischen Arbeitsschritten mit der Einheit *Minuten*. Alle Aspekte eines Produkts sind in Form von *Aspektmodellen* im *Aspect Model Catalog* mensch- und maschinenlesbar gespeichert. Die Informationen über die digitalen Zwillinge werden über APIs für Top-Level-Anwendungen wie die bereits genannten Insights-Module nutzbar gemacht. [5]

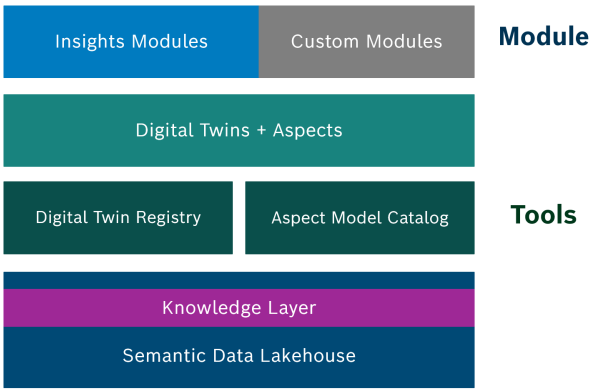


Abb. 1: vereinfachte, schematische Anordnung der Komponenten des Bosch Semantic Stack [2]

Zielsetzung der Arbeit

Die Bausteine oberhalb des Semantic Data Lakehouse sind als verteilte Anwendungen konzipiert. Da sie auf verschiedene Wege aufgerufen werden müssen, soll untersucht werden in welcher Form sich die Zugriffszeit auf benötigte Funktionen, die Erfüllung anwendungsübergreifender Aufgaben und die allgemeine Nutzererfahrung mit den Modulen und Tools durch eine zentrale Anlaufstelle und einen rollenbasierten Einstieg verbessert. Auf Basis der daraus abgeleiteten Anforderungen sollen *Low-Fidelity Prototypen* der benötigten Oberflächen entwickelt und validiert werden. Dafür wurde folgendes Mission Statement formuliert: „Der *Bosch Semantic Stack Workspace* ist ein zentraler Ort, der eine einheitliche Nutzeroberfläche für den Einstieg in alle Bosch Semantic Stack Module und Tools sowie relevante Informationen und Dokumentation bietet, um allen Personas ab dem Login über die gesamte User Journey hinweg nahtlose Interaktion zwischen allen Anwendungen zu ermöglichen, im Gegensatz zu einem offenen Marktplatz für Kundenlösungen“. Daraus ergibt sich die Anforderung an eine einmalige, modulübergreifende Registrierung und Verknüpfung mit dem jeweiligen Access-Management sowie die Bündelung der benötigten Tools und Module im Workspace, um eine einheitliche UX anhand einer durchgängigen Toolchain und einheitlichem Look and Feel zu erreichen.

Methodik

Die Entwicklung orientiert sich am menschzentrierten Gestaltungsprozess gemäß ISO 9241-210. Zu Beginn wird im Rahmen eines Kickoff-Workshops im *Question Zero*-Format gemeinsamer Konsens geschaffen, der die verschiedenen Perspektiven der Stakeholder berücksichtigt und die Hypothese zur zentralen Anlaufstelle konkretisiert. Die relevanten Nutzergruppen werden anschließend über eine *Stakeholder-Map* identifiziert. Nachfolgend werden qualitative Interviews über die Aufgabenbereiche der Nutzer sowie deren Interaktionen mit den verschiedenen Anwendungen des Semantic Stack durchgeführt. Anhand der identifizierten Nutzungshürden und Wünsche werden spezifische Anforderungen abgeleitet, auf deren Basis Wireframes der benötigten Benutzeroberflächen konzipiert werden. Darauf folgt die Evaluation der Gebrauchstauglichkeit der Entwürfe durch Nutzervertreter.

Ergebnisse

Zur Anforderungsermittlung wurden 15 Stakeholder befragt, die die Lösungen des Semantic Stack direkt und indirekt nutzen oder in ihrer Organisation einführen. Beteiligt waren Rollen aus den Bereichen Data/Knowledge Engineering, Solution Consulting, Business Development/Consultative Selling, Projektmanagement und Entwicklung.

Aktueller Stand

Die intensive Befragung der Nutzer sowie die gründliche Analyse dieser Daten bilden die Grundlage für die nachfolgende Ideation Phase. Basierend auf den abgeleiteten Anforderungen an den Workspace werden mit dem Protoyping-Tool „Balsamiq“ Low-Fidelity Prototypen der benötigten Oberflächen und deren Interaktionsmöglichkeiten visualisiert. Zudem werden alternative Gestaltungslösungen erprobt, wie es in der ISO 9241-210 vorgesehen ist. [3] Die *Mockups* werden von den zuvor identifizierten Nutzern validiert und anschließend an das Produktentwicklungsteam und professionelle UX-Designer übergeben, die sie zu High-Fidelity Prototypen weiterentwickeln, um den Weg für eine baldige Produktivsetzung zu ebnen.

Literatur und Abbildungen

[1] Jonas Busse, Bernhard Humm, Christoph Lübbert, Frank Moelter, Anatol Reibold, Matthias Rewald, et al. Was bedeutet eigentlich Ontologie? Ein Begriff aus der Philosophie im Licht verschiedener Disziplinen. *Informatik Spektrum*, 37:286, 2014.

[2] Eigene Darstellung.

[3] DIN-Normenausschuss Ergonomie Deutsches Institut für Normung. Ergonomie der Mensch-System-Interaktion - Teil 210: Menschzentrierte Gestaltung interaktiver Systeme (ISO 9241-210:2019). *DIN EN ISO 9241-210*, page 26, 2019.

[4] European Commission Directorate-General for Environment DG-ENV. Circular economy: New law on more sustainable, circular and safe batteries enters into force. https://environment.ec.europa.eu/news/new-law-more-sustainable-circular-and-safe-batteries-enters-force-2023-08-17_en?prefLang=de, 08 2023.

[5] Robert Bosch Manufacturing Solutions GmbH. About the Digital Twin Registry. https://docs.bosch-semantic-stack.com/registry/index.html, 05 2025.

[6] Robert Bosch Manufacturing Solutions GmbH. Bosch Semantic Stack Produktzentrierte digitale Transformation. https://www.bosch-connected-industry.com/de/de/portfolio/bosch-semantic-stack, 01 2025.

[7] Robert Bosch Manufacturing Solutions GmbH. Understand Aspect Models to reveal semantics of data. https://docs.bosch-semantic-stack.com/getting-started/understand-aspect-models.html, 04 2025.

[8] Ronald Hartwig, Sascha Wolter, and Martin Beschnitt. Usability & User Experience - Software näher zum Nutzer bringen. *Usability & User Experience Bitkom*, page 33, 2017.

[9] Jonas Rashedi. Was ist ein Datensilo? https://www.springerprofessional.de/datenmanagement/crm/was-ist-ein-datensilo-/18510004, 10 2020.

[10] Joseph Tsidulko. Was versteht man unter Datensilos? Warum sind sie problematisch? https://www.oracle.com/de/database/data-silos/, 01 2024.

Entwicklung eines RPA-basierten Workflows mit IDP-integration

Joel Schlossarek **Jürgen Koch**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma BMW AG, München

Einleitung

Prozessautomatisierung bezeichnet den gezielten Einsatz technologischer Systeme zur automatisierten Ausführung wiederkehrender, regelbasierter Tätigkeiten ohne direktes menschliches Eingreifen. Ziel ist es, betriebliche Abläufe effizienter zu gestalten, Fehlerquellen zu minimieren und Durchlaufzeiten signifikant zu reduzieren. Die zentralen Vorteile liegen in der Erhöhung der Produktivität, der Senkung operativer Kosten sowie in der Entlastung von Mitarbeitenden durch die Übernahme repetitiver Aufgaben.
Die technische Realisierung solcher Automatisierungsvorhaben scheitert jedoch häufig an den hohen Aufwänden für die Bereitstellung geeigneter Schnittstellen oder der notwendigen Neu- oder Umprogrammierung bestehender Anwendungssysteme [5]. Um diese Herausforderungen zu umgehen, bietet sich der Einsatz von Robotic Process Automation (RPA) an. RPA ermöglicht die Automatisierung von Geschäftsprozessen, ohne dass Änderungen an den bestehenden IT-Systemen oder Anwendungen erforderlich sind. Auf eine aufwendige Schnittstellenintegration kann dabei in der Regel verzichtet werden.

Robotic Process Automation

RPA bezeichnet eine Methode zur Automatisierung repetitiver, regelbasierter Geschäftsprozesse. Sie ermöglicht es Organisationen, strukturierte Aufgaben durch sogenannte Software-Roboter effizient und konsistent auszuführen. Der Begriff setzt sich aus den drei Komponenten Robotic, Process und Automation zusammen. Robotic verweist hierbei nicht auf physische Maschinen, sondern auf virtuelle, softwarebasierte Agenten. Process beschreibt den zugrunde liegenden Geschäftsablauf, der automatisiert werden soll, während Automation die Zielsetzung einer weitgehend autonomen Ausführung ohne manuelle Eingriffe betont. RPA trägt somit wesentlich zur Steigerung von Prozesseffizienz, Standardisierung und Fehlerreduktion bei [1].

Digitale Software-Roboter innerhalb von RPA-Systemen interagieren über die Benutzeroberfläche (englisch: User Interface) mit IT-Anwendungen, analog zum menschlichen Anwender [3]. Sie sind in der Lage, Mausbewegungen, Tastatureingaben sowie das Navigieren durch grafische Oberflächen zu simulieren und können so regelbasierte Arbeitsabläufe vollständig automatisiert ausführen. Aufgrund dieser Fähigkeit zur Nachahmung menschlicher Interaktionen werden solche Software-Roboter häufig als „virtuelle Mitarbeiter“ bezeichnet [4]. Ein wesentlicher Vorteil dieser Vorgehensweise besteht darin, dass keine Änderungen an den bestehenden IT-Systemen oder deren zugrunde liegender Infrastruktur erforderlich sind [3].

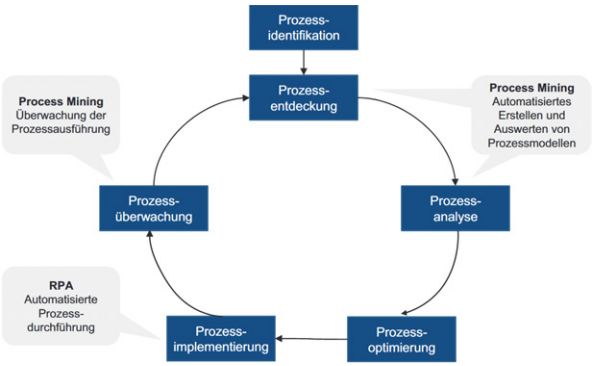


Abb. 1: RPA-Lifecycle [2]

RPA folgt einem zyklischen Lebenszyklus, der sich an den Prinzipien des Geschäftsprozessmanagements orientiert, um Prozesse effizient zu automatisieren. Der Zyklus beginnt mit der Prozessidentifikation, in der geeignete Automatisierungskandidaten erkannt werden. Darauf folgt die Prozessentdeckung, bei der die Ist-Prozesse detailliert dokumentiert und analysiert werden. In der anschließenden Prozessanalyse erfolgt eine Bewertung hinsichtlich technischer Umsetzbarkeit und wirtschaftlichem Potenzial. Die Prozessoptimierung dient der gezielten Verbesserung des Ablaufs, bevor die Prozessimplementierung in Form eines RPA-Workflows

erfolgt. Nach erfolgreichem Testing wird der Bot produktiv eingesetzt. Die Prozessüberwachung stellt durch Monitoring der Ausführung und entsprechendes Change Management sicher, dass die Automatisierung stabil und anpassungsfähig bleibt [2].

Intelligent Document Processing

Intelligent Document Processing (IDP) basiert auf dem Einsatz von Künstlicher Intelligenz (KI) und Machine Learning (ML), um Informationen aus verschiedenen Dokumentenarten zu erfassen, zu analysieren und in eine verwertbare Form zu bringen. Unabhängig davon, ob es sich um klar strukturierte, teilstrukturierte oder unstrukturierte Dokumente handelt, ermöglicht IDP eine inhaltliche Erschließung, die dem menschlichen Interpretationsvermögen nahekommt [6].

Ziel

Im Rahmen dieser Arbeit soll ein unternehmensinterner Bestellprozess bei der BMW AG durch den kombinierten Einsatz von RPA und IDP automatisiert werden. Ziel ist es, den Anwenderinnen und Anwendern eine Funktionalität bereitzustellen, mit der sich Dokumente hochladen lassen, deren Inhalte anschließend automatisiert analysiert, interpretiert und weiterverarbeitet werden. Basierend auf den extrahierten Informationen soll eine adäquate Bestellung generiert und an die entsprechenden Beschaffungssysteme übermittelt werden.

Zur Realisierung der Prozessautomatisierung werden verschiedene Softwarelösungen eingesetzt. UiPath wird dabei als RPA-Plattform verwendet, während der AI Builder in Kombination mit Microsoft Power Apps die Funktionalitäten des IDP übernimmt. Die zentrale Datenablage erfolgt über Microsoft SharePoint.

Herausforderungen

Im Rahmen der Entwicklung der Prozessautomatisierung waren bereits eine Reihe organisatorischer und technischer Herausforderungen zu bewältigen. Dazu zählen unter anderem die Durchführung umfangreicher Schulungen, die Beantragung erforderlicher Softwarelizenzen sowie die Berücksichtigung und Einhaltung unternehmensspezifischer Compliance-Richtlinien.
Eine besondere Komplexität ergibt sich aus der hohen Varianz der zu automatisierenden Bestellprozesse, die von der Beschaffung einfacher Büromaterialien über die Buchung von Tickets bis hin zur Bestellung von Einzelteilen reicht. Diese Prozesse erfordern jeweils die Anbindung unterschiedlicher IT-Systeme sowie die Abbildung individueller Prozesslogiken.
Eine zentrale Anforderung an die zu entwickelnde Automatisierungslösung besteht darin, sowohl die spezifischen Beschaffungsrichtlinien der BMW AG zu berücksichtigen als auch eine reibungslose Integration in die bestehende IT-Systemlandschaft sicherzustellen.

Literatur und Abbildungen

[1] Vivek Bhardwaj, Shveta Yadav, Navjeet Kaur, and Darpan Anand. The Future of Work: Robotic Process Automation and its Role in Shaping Tomorrow's Business Landscape. *SN Computer Science*, 6:111, 2025.

[2] Carsten Feldmann. *Praxishandbuch Robotic Process Automation (RPA): Von der Prozessanalyse bis zum Betrieb*. Springer Fachmedien Wiesbaden, 2022.

[3] Christian Gärtner. *Smart HRM: Digitale Tools für die Personalarbeit*. Springer Fachmedien Wiesbaden, 2 edition, 2024.

[4] Ennis Gündoğan. *Robotic Process Automation (RPA) im Desktop-Publishing: Softwaregestützte Automatisierung von Artwork-Prozessen*. Springer Fachmedien Wiesbaden, 2025.

[5] Mario Richard Smeets, Ralf Jürgen Ostendorf, and Andreas Freßmann. *Robotic Process Automation im Einsatz: Strategische Ausrichtung – praktische Umsetzung – revisionssichere Implementierung*. Springer Fachmedien Wiesbaden, 2023.

[6] Maxime Vermeir. What is Intelligent Document Processing: Benefits, Use Cases. https://www.abbyy.com/blog/intelligent-document-processing/, 12 2021.

Prototyping eines ML Modells zur Berechnungszeitvorhersage für SAT Solver basierte Services

Simon Schoppe **Dirk Hesse**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma BMW AG, München

Einleitung

In der heutigen digitalen Ära sind effiziente Problemlösungsmethoden von entscheidender Bedeutung. In Bereichen, die eine komplexe Entscheidungsfindung benötigen ist diese Effizienz besonders gefordert. Ein prominentes Beispiel hierfür sind *Satisfiability Solver* (SAT-Solver). SAT-Solver lösen Erfüllbarkeitsprobleme, auch genannte SAT-Probleme, die aus logischen Formeln bestehen. Die Effizienz der SAT-Solver hängt jedoch entscheidend von der Komplexität des zu lösenden Problems ab. Dies führt zu Variationen in der Berechnungszeit für die Lösung der einzelnen logischen Formeln. Eine bessere Planung der Rechenressourcen des Services würde dem Benutzer, also dem Anfragesteller, eine schnellere Antwort liefern und somit die Benutzererfahrung optimieren. Um dieses Ziel zu erreichen, soll ein präzises Modell die Berechnungszeit der Anfragen vorhersagen. In dieser Arbeit wird ein erster Prototyp eines *Machine Learning* Modell (ML-Modell) entwickelt, das eine Vorhersage der Berechnungszeit treffen kann. Hierbei wird sowohl die Kategorie des Modells als auch das eigentliche Modell ausgewählt. Darüber hinaus werden der Trainingsprozess und der Prozess zur Verbesserung der Genauigkeit beschrieben.

Problemstellung

Im Betriebsumfeld des Praxispartners ist ein SAT-Solver im Einsatz, der sich um verschiedene Anfragen kümmert. Die Anfragen werden von Recheneinheiten, sogenannten Pods, bearbeitet. Es sind zwischen 50 und 300 dieser Pods im Einsatz und benötigen je nach Anfragetyp und -komplexität länger diese zu bearbeiten. Die Variation der Bearbeitungszeit befindet sich zwischen wenigen Millisekunden bis hin zu mehreren Minuten. Im aktuellen Betriebsumfeld werden die Anfragen nach dem Prinzip *First Come First Serve* bearbeitet. Da die Länge der Bearbeitungszeit für die Zuweisung der Anfragen zu den einzelnen Pods nicht bekannt ist, kann es zu einem Stau der Anfragen führen. Wenn beispielsweise alle Pods mit Anfragen belegt sind, die eine Bearbeitungszeit von mehreren Minuten benötigen, können sich dadurch Anfragen in der Warteschlange häufen. Daraus resultiert eine längere Wartezeit für die Anfragesteller, da alle Pods ausgelastet sind und anstehende Anfragen nicht bearbeitet werden können. Wie in der Einleitung bereits erwähnt, soll ein ML-Modell zu einer verkürzten Antwortzeit beitragen und dementsprechend bei einer Optimierung unterstützen.

Methodik

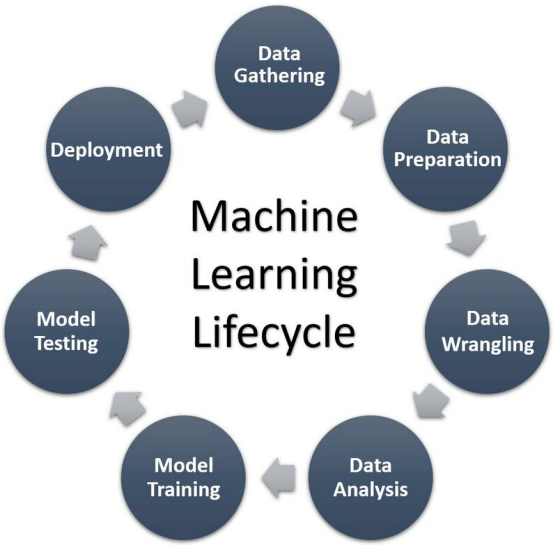


Abb. 1: Life Cycle über die Entwicklung eines ML Modells [1]

In Abbildung 1 sind die einzelnen Schritte, um ein ML-Modell zu entwickeln, dargestellt. Um das in

Kapitel Problemstellung beschriebene Problem zu lösen, wurden die Anforderungen an das ML-Modell mit dem Praxispartner ermittelt und erste Entscheidungsoptionen gefunden. Die erste Entscheidungsoption beinhaltet die Festlegung auf die Kategorie des ML-Modells. Sowohl *Classification* als auch *Regression* sind für diese Problemstellung möglich und vom Praxispartner nicht ausgeschlossen. In Abbildung 2 wird diese Entscheidung unter „Predict Numeric“ im mittigen ovalen Feld dargestellt. Bei einem Nein wird sich für die Kategorie *Classification* entschieden und bei einem Ja für die Kategorie *Regression*. Der nächste Schritt ist die Sammlung von möglichen Daten, also dem *Data Gathering* aus Abbildung 1. Hierzu wurde bereits eine Speicherung aller Anfragen mit Bearbeitungszeit an den SAT-Solver vorgenommen. Bei den nachfolgenden drei Schritten, also der *Data Preparation*, dem *Data Wrangling* und der *Data Analysis* aus Abblildung 1, werden die Daten auf Vollständigkeit geprüft, ausgewertet und eventuell gefiltert. Dies führt dazu, dass die relevanten Parameter aus den Anfragen extrahiert werden und Zusammenhänge zwischen den Anfragemerkmalen und der Bearbeitungszeit gefunden wird. Eine Unterstützung bei der Findung eines passenden Modells gibt der Entscheidungspfad von Abbildung 2. Durch die erste Entscheidung kann sich innerhalb der festgelegten Kategorie weiter orientiert werden und das Modell danach ausgesucht werden.

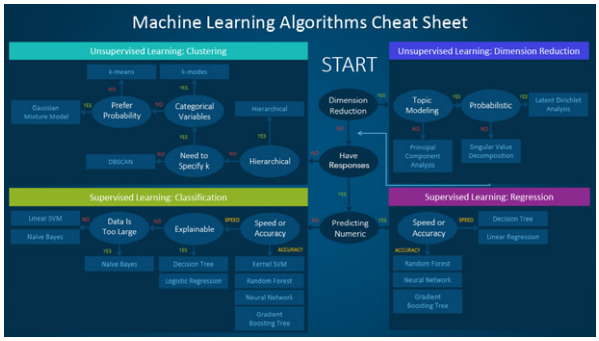


Abb. 2: Entscheidungsbaum für die Auswahl eines geeigneten ML Modells [2]

Im weiteren Verlauf der Thesis wird beschrieben, wie das ausgewählte Modell mit den ausgewählten Daten trainiert wird und welche Genauigkeit das Modell erreicht. Um die Genauigkeit des Modells zu verbessern, kann aus den gewonnenen Resultaten noch ein *Fine Tuning*, also eine Feinabstimmung der Trainingsparameter, durchgeführt werden. Eine weitere Möglichkeit, um die Genauigkeit zu erhöhen ist die *Cross Validation*. Wie in Abbildung 3 dargestellt, wird der Trainingsdatensatz in k Anteile aufgeteilt. Einer dieser Aufteilungen wird pro Split als Validierungsdatensatz hergenommen und damit das Training des Modells bewertet. Durch dieses Vorgehen kann sowohl verhindert werden das der Trainingsdatensatz verkleinert wird als auch das ein sogenanntes *Overfitting* stattfindet. [3]

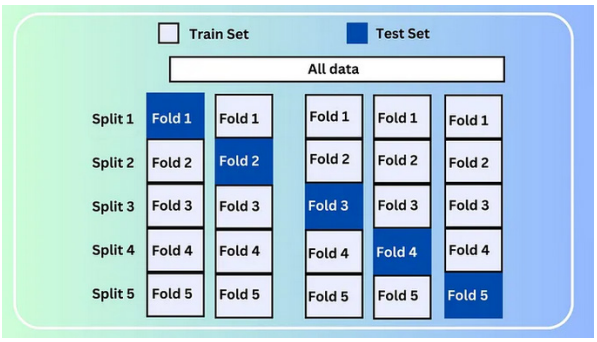


Abb. 3: Grundprinzip von Cross Validation [3]

Zusammenfassung

Die Bachelorarbeit befasst sich mit der Entwicklung eines Prototypens für ein ML-Modell. Dieses Modell soll die Bearbeitungszeit der Anfragen vorherzusagen, um die Zuweisung der Pods zu optimieren. Für die Vorhersage der Berechnungszeit wird anfangs die vorliegenden Daten von allen Anfragen ausgewertet. Durch die Auswertung können wichtige Anfragemerkmale, die die Bearbeitungszeit stark beeinflussen, herausgefiltert werden. Mithilfe dieser Erkenntnisse kann sich für die optimale Kategorie des Modells und damit auch für das passende Modell entschieden werden. Im nächsten Schritt wird das ausgewählte ML-Modell trainiert. Um die Genauigkeit des Modells zu verbessern kann ein *Fine Tuning*, also eine Feinabstimmung, durchgeführt werden. Dieses kann durch die Feinabstimmung der Trainingsparameter realisiert werden. Eine andere Möglichkeit besteht mit der *Cross Validation*, da durch Training mit veränderten Trainings- und Validierungsdatensatz das beste Ergebnis benutzt werden kann.

Literatur und Abbildungen

[1] Moez Krichen et al. Are Formal Methods Applicable To Machine Learning And Artificial Intelligence? https://www.researchgate.net/publication/362485323_Are_Formal_Methods_Applicable_To_Machine_Learning_And_Artificial_Intelligence, 05 2022.

[2] Hui Li. Which machine learning algorithm should I use? https://blogs.sas.com/content/subconsciousmusings/2020/12/09/machine-learning-algorithm-use/, 12 2020.

[3] Balaji Nalawade. The Essential Guide to K-Fold Cross-Validation in Machine Learning. https://medium.com/@bididudy/the-essential-guide-to-k-fold-cross-validation-in-machine-learning-2bcb58c50578, 05 2024.

Multi-Agenten-KI-Systeme für Business Process Improvement: Eine strukturierte Literaturrecherche zu Systemanforderungen

Nico Schurr **Manfred Schoch**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

Angesichts einer Geschäftswelt, die sich immer mehr dynamisch und wettbewerbsintensiv gestaltet, stehen Unternehmen vor der Aufgabe, ihre Geschäftsprozesse fortwährend zu optimieren, um Effizienz, Qualität und Innovationskraft zu erhöhen. Obwohl Business Process Improvement (BPI) als bewährter Ansatz gilt, sind die derzeitigen Verfahren häufig stark manuell, ressourcenaufwendig und nur begrenzt anpassungsfähig. Vor allem die kreative Neugestaltung und die multidimensionale Optimierung von Prozessen sind bislang kaum automatisierbar. In dieser Arbeit wird untersucht, inwiefern Multi-Agenten-KI-Systeme eine automatisierte Unterstützung von BPI leisten können. Ziel der Arbeit ist es daher, die wesentlichen Systemanforderungen zu bestimmen, die erfüllt sein müssen, damit ein solches System als intelligente, kontextsensitive und kreative Unterstützung im BPI-Prozess fungieren kann.

Theoretischer Hintergrund

Business Process Improvement (BPI) umfasst die systematische Analyse, Modellierung, Bewertung und Neugestaltung von Geschäftsprozessen, um Effizienz, Kundenorientierung oder Innovationsfähigkeit zu steigern. [4] Je nach strategischen Ziele und Ausgangslage des Unternehmens kann es sich dabei um inkrementelle oder radikale Verbesserungen handeln. Momentan konzentriert sich die Automatisierung von BPI vor allem auf die Untersuchung bestehender Abläufe. Mit Verfahren wie dem Process Mining können reale Prozessabläufe aus Event Logs extrahiert, visualisiert und bewertet werden. [8] Auf dieser Grundlage werden Verfahren der Künstlichen Intelligenz (KI) zunehmend angewendet, beispielsweise zur Klassifizierung von Anomalien oder zur Unterstützung von Entscheidungen in bestimmten Prozesssituationen. [5] Trotzdem sind viele BPI-Aktivitäten, vor allem die kreative Neugestaltung und multidimensionale Bewertung von Soll-Prozessen, bisher manuell und erfahrungsgetrieben. [9] Die Automatisierung von BPI in diesem Bereich verspricht eine verbesserte Skalierbarkeit, reduzierte Umsetzungszeiten und objektivere Entscheidungen. Trotzdem gibt es beträchtliche Herausforderungen. Kreative und strategische Tätigkeiten, wie das Ausarbeiten innovativer Prozessvarianten oder das Abwägen gegensätzlicher Zielsetzungen, sind nur schwer mit herkömmlichen Algorithmen darzustellen. [1] An dieser Stelle kommt der Einsatz Künstlicher Intelligenz in der Form von Multi-Agenten-KI-Systemen ins Spiel. Künstliche Intelligenz (KI) bezieht sich auf die Fähigkeit von technischen Systemen, Aufgaben zu erfüllen, die normalerweise menschliche Intelligenz erfordern, wie etwa Problemlösung, Mustererkennung oder Entscheidungsfindung. [6] Generative KI (GenAI) stellt eine besonders aktuelle Form dar und kann neue Inhalte wie Texte, Prozesse oder Designs basierend auf umfangreichen Trainingsdaten erstellen. [2] Da KI-Systeme grundsätzlich dazu bestimmt sind, Informationen zu verarbeiten und kluge Entscheidungen zu fällen, ist die Frage von Interesse, wie diese Fähigkeiten in ein interaktives System umgesetzt werden können, das handlungsfähig ist. Hier greift das Konzept des Agenten. Ein Agent verwendet KI, um neben der Analyse eigenständig in seiner Umgebung zu handeln, zu lernen und mit anderen Entitäten zu interagieren. In der Informatik versteht man unter einem Agenten ein autonom agierendes, softwarebasiertes System, das innerhalb einer Umgebung operiert, Ziele verfolgt und dabei auf Wahrnehmung, Entscheidungslogik und Handlung zurückgreift. [10] Das Konzept des Multi-Agenten-Systems (MAS) setzt genau hier an. Ein MAS setzt sich aus einer Gruppe von Agenten zusammen, die in einer gemeinsamen Umgebung nebeneinander existieren und miteinander kommunizieren, kooperieren oder konkurrieren, um entweder gemeinsame oder individuelle Ziele zu verfolgen. [10] Ein vereinfachtes

beispielhaftes MAS könnte im BPI-Kontext wie in Abbildung 1 dargestellt aussehen.

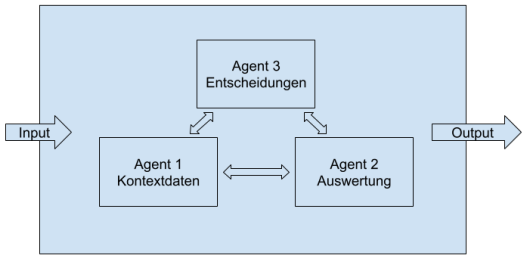


Abb. 1: MAS für Business Process Improvement [3]

Forschungsmethode

Zur Durchführung wird eine strukturierte Literaturrecherche im Rahmen der eDSR-Methode durchgeführt. Die wissenschaftliche Forschungsmethode Design Science Research (DSR) fokussiert sich auf die Entwicklung und Bewertung innovativer Artefakte, um reale Probleme in Organisationen oder Technologien zu lösen. Anders als bei rein erklärender Forschung liegt der Schwerpunkt bei DSR auf dem Gestalten (Design) und Testen (Evaluation) neuer Lösungen. Echeloned Design Science Research (eDSR)-Methodik erweitert traditionelle DSR-Modelle, die auf Phasen und Iterationen basieren, um eine zusätzliche Dimension.

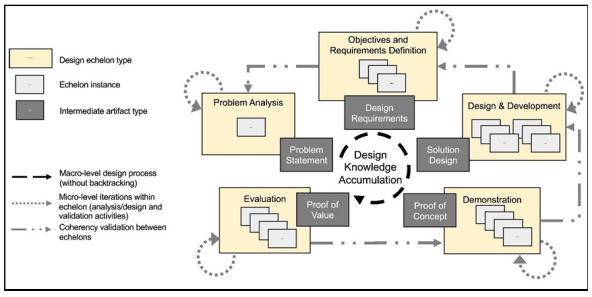


Abb. 2: eDSR Metamodell [7]

Wie in Abbildung 1 zu sehen, organisiert das Metamodell jeden dieser fünf Echelon-Typen (Problem Analysis, Objectives ß Requirements, Design ß Development, Demonstration und Evaluation) um ein bestimmtes Zwischenartefakt und verknüpft sie mit geeigneten Analyse-, Design- und Validierungsaktivitäten.

Ergebnisse

Zur Umsetzung eines entsprechenden Multi-Agenten-KI-Systems müssen folgende Faktoren berücksichtigt werden:

1. Berücksichtigung von Kontextfaktoren
2. Implementierung einer passenden Verarbeitungslogik
3. Festlegung der Datenressourcen
4. Technologische Infrastruktur
5. Menschliche Fähigkeiten
6. Visualisierung
7. Organisatorische Rahmenbedingungen
8. Definition der Systemparameter
9. Evaluierung/Monitoring

Ausblick

Diese Arbeit bietet eine fundierte Übersicht über funktionale Systemanforderungen bei der Konzeption agentenbasierter BPI-Systeme und verbindet damit verschiedene bislang getrennte Forschungsrichtungen, wie dem Business Process Management, der KI-Forschung und der Agentenarchitektur. Viele der identifizierten Anforderungen sind technologieoffen formuliert und erfordern eine weiterführende technische Spezifikation. Diese sollten in prototypische Architekturen integriert und in realen Unternehmensumgebungen evaluiert werden, um die zukünftige Arbeit zu verbessern.

Literatur und Abbildungen

[1] P. Afflerbach et al. Design it like Darwin - A value-based application of evolutionary algorithms for proper and unambiguous business process redesign. *Inf Syst Front*, 19, 2017.

[2] R. Bommasani. On the opportunities and risks of foundation models. *arXiv preprint*, 2022.

[3] Eigene Darstellung.

[4] TH. Davenport. *Process innovation: reengineering work through information technology.* Harvard Business Review Press, 1993.

[5] J. Ghattas et al. Improving business process decision making based on past experience. *Decision Support Systems*, 59, 2014.

[6] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Pearson, 2020.

[7] Tuure Tuunanen et al. Dealing with Complexity in Design Science Research: A Methodology Using Design Echelons. *MIS Quarterly*, 2024.

[8] W. van der Aalst. *Process Mining: Data Science in Action*. Springer, 2016.

[9] M. Voigt et al. Comprehensive support for creativity-intensive processes: An explanatory information system design theory. *Business ß Information Systems Engineering*, 5, 2013.

[10] M. Wooldridge. *An Introduction to MultiAgent Systems*. Wiley, 2009.

KI-gestützte Content-Generierung in Webanwendungen: Architektur und Umsetzung einer Backend-Erweiterung zur automatisierten Erstellung von Soft-Skill-Trainingsinhalten

Connor Schwab **Jörg Nitzsche**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Savvi Learning GmbH, Kornwestheim

Einleitung & Motivation

Die fortschreitende Entwicklung generativer KI-Technologien eröffnet neue Möglichkeiten zur automatisierten Erstellung digitaler Lerninhalte. Im Rahmen eines Praxisprojekts bei der Savvi Learning GmbH wurde im Zuge dieser Bachelorarbeit untersucht, wie sich eine bestehende Webanwendung architektonisch um eine skalierbare KI-Komponente erweitern lässt, um strukturierte Soft-Skill-Trainingsinhalte automatisiert zu erzeugen.

Ziel der Arbeit ist die Integration generativer KI, konkret gesagt OpenAI Assistants mit strukturierter JSON-Ausgabe, in eine bestehende Webarchitektur, die auf einem Django-Backend basiert. Die zentrale Forschungsfrage lautet:

„Wie lässt sich eine bestehende Backend-Architektur einer Webanwendung um eine skalierbare KI-Komponente zur automatisierten Generierung strukturierter Soft-Skill-Lerninhalte erweitern?“

Der Fokus liegt auf der Entwicklung einer robusten, modularen und skalierbaren Systemarchitektur, die die Integration von asynchronen KI-Prozessen erlaubt und gleichzeitig eine hohe Wartbarkeit und Nachvollziehbarkeit im Betrieb sicherstellt.

Systemarchitektur & Technologie-Stack

Die gewählte Architektur basiert auf modernen, containerisierbaren Webtechnologien, die eine klare Trennung von Zuständigkeiten und eine skalierbare Verarbeitung sicherstellen.

Als relationale Datenbank dient PostgreSQL, die sich durch Transaktionssicherheit und die Unterstützung strukturierter Datenmodelle auszeichnet. Das zentrale Backend-Framework Django ermöglicht eine strukturierte Entwicklung nach dem MVT-Muster und stellt mit Komponenten wie ORM, Routing und Admin-Oberfläche eine stabile Grundlage für eine wartbare Backend-Architektur bereit.

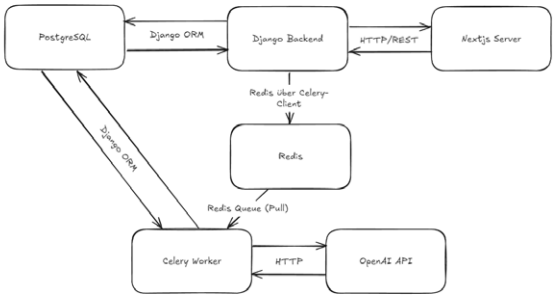


Abb. 1: Architektur Diagramm [2]

Für die Verarbeitung langlaufender oder rechenintensiver Prozesse – etwa bei der Nutzung generativer KI – wird das Task-Queue-System Celery verwendet, das sich für asynchrone Workflows in Python-basierten Webanwendungen etabliert hat [4]. Die Aufgaben werden dabei über Redis als Message-Broker verteilt, was eine lose Kopplung zwischen Webserver und Worker-Prozessen ermöglicht [3]. Diese Architektur trennt zeitkritische Benutzerinteraktionen von ressourcenintensiven Aufgaben wie KI-Anfragen und verbessert damit sowohl die Skalierbarkeit als auch die Reaktionsfähigkeit des Systems.

Die Anbindung an die OpenAI Assistant API erfolgt durch die Worker, die vordefinierte Prompts mitsamt strukturierter JSON-Ausgabe übermitteln. Die Rückgabe der Ergebnisse in strukturierter Form erleichtert die automatisierte Weiterverarbeitung im Backend. [1]

Die Frontend- und Serverschnittstelle bildet Next.js, das sowohl clientseitige Komponenten als auch serverseitige API-Endpunkte bereitstellt. Es übernimmt die Vermittlung zwischen Benutzerinteraktion und Backend-Logik und unterstützt zugleich reaktive UI-

Elemente sowie serverseitig gerenderte Inhalte für eine performante Darstellung.
Diese Komponenten werden in einem modularen Architekturmodell zusammengeführt (vgl. Abbildung 1), das sowohl horizontale Skalierung als auch entkoppelte Entwicklung einzelner Subsysteme unterstützt.

Konzeptioneller Ablauf

Der Ablauf des KI-basierten Content-Generierungsprozesses gestaltet sich wie folgt:
Die Interaktion beginnt mit einem Nutzerkommando im Frontend, das über den Next.js-Client an den zugehörigen Server weitergeleitet wird. Dieser sendet eine API-Request an das Django-Backend, wo ein neues Statusobjekt erstellt und eine zugehörige ‘task_uuid‘ generiert wird. Diese UUID dient der späteren Identifikation und Fortschrittsverfolgung des Generierungsprozesses und wird unmittelbar an den Client zurückgegeben.
Parallel wird im Backend ein asynchroner Task über Celery gestartet. Dieser ruft die OpenAI Assistant API auf, wobei ein vordefinierter Prompt samt strukturellem Output-Schema übermittelt wird. Die OpenAI-API verarbeitet die Anfrage und liefert eine strukturierte JSON-Antwort zurück, die mehrere Suggestion-Objekte enthält – also alternative inhaltliche Vorschläge für die angeforderte Lerneinheit. Diese Vorschläge werden im Backend persistiert und dem jeweiligen Statusobjekt zugeordnet.
Sobald die Verarbeitung abgeschlossen ist, erkennt ein in Django implementierter Signal-Handler die Statusaktualisierung. Dieser stößt eine Kommunikation per Webhook an den Next.js-Server an, wodurch das Frontend über den Abschluss der Generierung informiert werden kann. Die Nutzeroberfläche aktualisiert sich daraufhin automatisch und zeigt die generierten Inhalte zur Auswahl an.
Der gesamte Ablauf ist so konzipiert, dass er nichtblockierend, modular und durchgehend nachvollziehbar ist. Die Kombination aus synchronem Feedback (durch Rückgabe der UUID) und asynchroner Generierung (über Worker und Webhooks) stellt sicher, dass Nutzerinteraktionen nicht durch langlaufende Prozesse verzögert werden und der Status jederzeit rekonstruierbar bleibt.

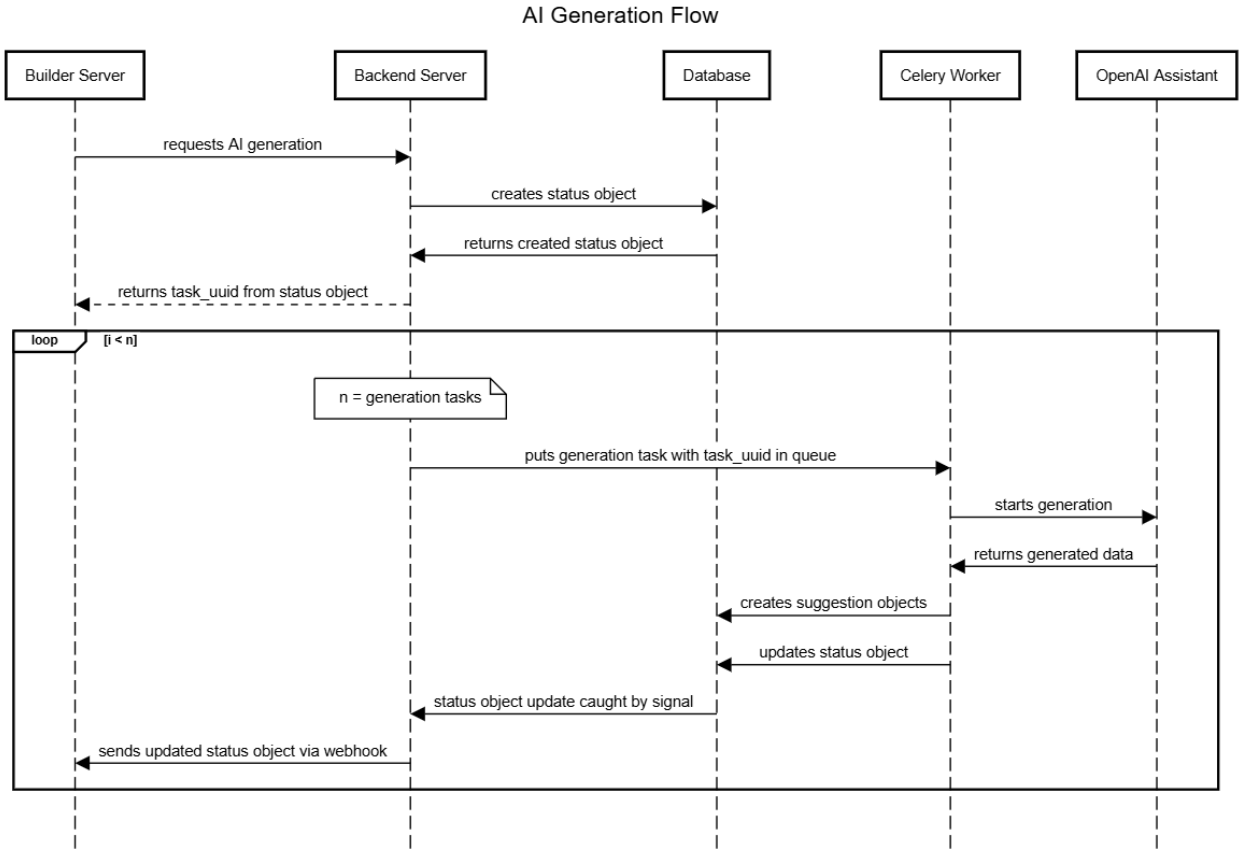


Abb. 2: Sequenz-Diagramm des AI-Generation-Flows [2]

Ausblick: Implementierungsstrategie und Schwerpunkte

Die Implementierung legt besonderen Wert auf eine modulare und dynamisch erweiterbare Systemarchitektur. Neue KI-Komponenten oder Prompt-Typen lassen sich dank eines pluginartigen Aufbaus problemlos integrieren, ohne bestehende Funktionalitäten zu beeinträchtigen. Um die asynchronen Prozesse nachvollziehbar zu gestalten, wird ein zentrales Logging-System auf Basis von Promtail, Loki und Grafana eingesetzt. Dieses ermöglicht die strukturierte Protokollierung und Visualisierung der Workeraktivitäten sowie die Analyse potenzieller Fehlverläufe im Betrieb.

Darüber hinaus wird die Testbarkeit des Systems durch gezielte Unit-Tests sichergestellt, insbesondere für die Backend-Logik, die API-Schnittstellen sowie die signalbasierten Ereignisflüsse. Ein weiterer zentraler Aspekt ist das Prompt Engineering. Die Qualität der von der KI generierten Inhalte hängt maßgeblich von der Gestaltung der Prompts und der bereitgestellten Kontexte ab. Dieser Teil des Systems wird iterativ optimiert, um sowohl die Struktur als auch die inhaltliche Konsistenz der Vorschläge zu gewährleisten.

Literatur und Abbildungen

[1] Robert Craigie. OpenAI API – Python SDK (openai-python). Offizielle GitHub-Dokumentation. https://github.com/openai/openai-python, 2023.

[2] Eigene Darstellung.

[3] Ask Solem. Backends and Brokers. https://docs.celeryq.dev/en/stable/getting-started/backends-and-brokers/, 2023.

[4] Giancarlo Zaccone. *Python Parallel Programming Cookbook*. Packt Publishing, 2015.

Lift-and-Shift-Migration zu einer Serverless-Architektur: Analyse, Prototyping und Evaluation anhand einer Verkaufsanwendung

Arbresha Selimi **Dennis Grewe**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

Cloud Computing entwickelt sich kontinuierlich weiter und schafft somit neue Möglichkeiten zur Optimierung bestehender Prozesse. [5] Besonders das Serverless-Paradigma hat seit der Einführung von AWS Lambda im Jahr 2014 deutlich an Bedeutung gewonnen. Durch die vollständige Übernahme des Servermanagements durch den Cloud-Anbieter gilt Serverless Computing heute als vielversprechender Architekturansatz, der sowohl in der Forschung als auch in der Industrie zunehmend Beachtung findet. [5]
Laut der Statistik „Adoption of IT Trends in North America and Europe 2025“ von Spiceworks setzen bereits 32 % der Unternehmen in Europa und Nordamerika Serverless-Technologien ein, und weitere 18 % planen deren Einführung innerhalb der nächsten zwei Jahre. [2] Viele bedeutende Unternehmen sind daran interessiert, geeignete Vertriebsanwendungen in eine Serverless-Architektur zu überführen, um von Vorteilen wie reduzierten Betriebsaufwänden zu profitieren. Gleichzeitig bringt der Serverless-Ansatz jedoch neue Herausforderungen in der Entwicklung, im Betrieb und in der Überwachung mit sich, die analysiert und bewertet werden müssen.

Zielsetzung

Ziel der Thesis ist es, zu untersuchen, inwiefern sich Serverless Computing als geeigneter Ansatz für unternehmerische Anwendungen eignet. Im Mittelpunkt steht dabei die Frage, welche Auswirkungen die Migration einer containerbasierten Anwendung auf zentrale Kennzahlen wie Betriebskosten, Skalierbarkeit und Performance hat und welche Herausforderungen dabei zu bewältigen sind. Zu diesem Zweck wird in einer Fallstudie eine bestehende Spring-Boot-Anwendung mittels Lift-and-Shift-Ansatz in eine Serverless-Architektur überführt.

Der direkte Vergleich der containerbasierten und der Serverless-Architektur erfolgt anschließend anhand gezielter Messungen. Anhand der gewonnenen Ergebnisse wird bewertet, unter welchen Voraussetzungen der Einsatz von Serverless Computing wirtschaftlich und technisch vorteilhaft ist, insbesondere im Kontext der betrachteten Anwendung.

Serverless Computing

Aufbauend auf der Analyse von Samuel Kounev et al. [3] lässt sich Serverless Computing als ein Cloud-Computing-Paradigma beschreiben, das die Entwicklung, Bereitstellung und Ausführung von Anwendungen oder Komponenten ermöglicht, ohne dass Entwickler Server oder Infrastrukturen verwalten müssen. Der Cloud-Anbieter übernimmt alle operativen Aufgaben wie Fehlertoleranz, elastische Skalierung und Ressourcenmanagement, um den Anforderungen der Anwendung gerecht zu werden. Die Abrechnung erfolgt nutzungsbasiert und feingranular, sodass nur tatsächlich beanspruchte Ressourcen bezahlt werden müssen.
Im Rahmen dieser Arbeit wird Serverless Computing anhand des Modells Function-as-a-Service realisiert.

Function-as-a-Service

Function-as-a-Service ist ein Modell zur Implementierung von Serverless Computing, bei dem Anwendungen aus einzelnen, ereignisgesteuerten Funktionen bestehen. Diese Funktionen sind zustandslos und werden vom Cloud-Anbieter nur dann instanziiert, wenn sie tatsächlich benötigt werden. Nach der erfolgreichen Ausführung werden sie automatisch wieder beendet. Dadurch werden Ressourcen effizient genutzt und eine automatische Skalierung erfolgt bedarfsgerecht. [4]
Im Gegensatz zu Platform-as-a-Service, bei dem in der Regel dauerhaft laufende Instanzen betrieben

werden müssen, entfallen bei Function-as-a-Service alle Hintergrundprozesse. Abgerechnet wird ausschließlich die tatsächliche Ausführungszeit, da bei Inaktivität keine Kosten entstehen. [4]

Fallbeispiel

Bei der zu migrierenden Anwendung handelt es sich um ein containerisiertes Spring-Boot-Backend zur Verarbeitung von Ereignissen. Es empfängt drei unterschiedliche Ereignistypen über Apache Kafka, einer Plattform für das fortlaufende Streamen von Ereignisdaten. Die Anwendung verarbeitet diese Ereignisse entsprechend und speichert sie in einer PostgreSQL-Datenbank.
Anschließend werden die verarbeiteten Ereignisse erneut an Kafka übermittelt, damit sie von weiteren externen Systemen verarbeitet werden können. Um die Konsistenz zwischen Datenbank und Ereignisübertragung sicherzustellen, wird das Transactional-Outbox-Pattern verwendet, das in Abbildung 1 dargestellt ist.

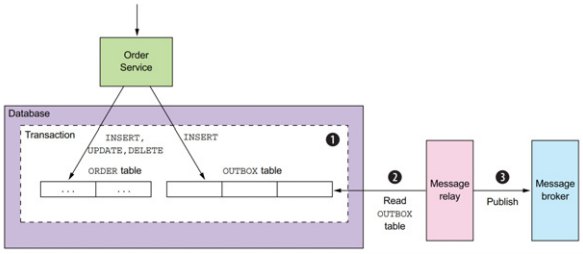


Abb. 1: Beispielhafter Ablauf des Transactional Outbox Patterns [6]

Im ersten Schritt werden innerhalb einer Transaktion sowohl die fachlichen Änderungen in der entsprechenden Tabelle als auch ein Event-Eintrag in der Outbox-Tabelle gespeichert. Im zweiten Schritt liest ein separater Prozess regelmäßig die Einträge aus der Outbox-Tabelle aus und sendet im letzten Schritt die enthaltenen Ereignisse an den Message-Broker. Dadurch wird verhindert, dass Ereignisse verloren gehen oder doppelt versendet werden. [6]
Im aktuellen Zustand läuft die Anwendung in Containern und erfordert dadurch eine dauerhafte Bereitstellung, die kontinuierliche Kosten verursacht. Da sie jedoch nur bei eingehenden Ereignissen aktiv ist, eignet sie sich besonders für eine Serverless-Architektur, bei der in inaktiven Phasen keine Betriebskosten entstehen.

Konzept

Um die Auswirkungen einer serverlosen Architektur bewerten zu können, soll auf Basis der bestehenden Anwendung schrittweise ein Prototyp nach dem Lift-and-Shift-Ansatz entwickelt werden. Die fachliche Logik bleibt dabei unverändert, während Architektur und Ausführungsumgebung angepasst werden. [7] So lassen sich die Effekte der Migration isoliert analysieren. Die prototypische Umsetzung soll die zentralen Komponenten als einzelne Cloud-Funktionen abbilden, wobei jede Funktion für einen bestimmten Ereignistyp zuständig ist. Zur Veranschaulichung dessen dient ein technologieunabhängiges Architekturdiagramm (vgl. Abbildung 2). Auch das Transactional-Outbox-Pattern soll weiterhin übernommen werden. Eine zusätzliche Funktion übernimmt die Verarbeitung der Outbox-Tabelle und den Versand der Ereignisse.

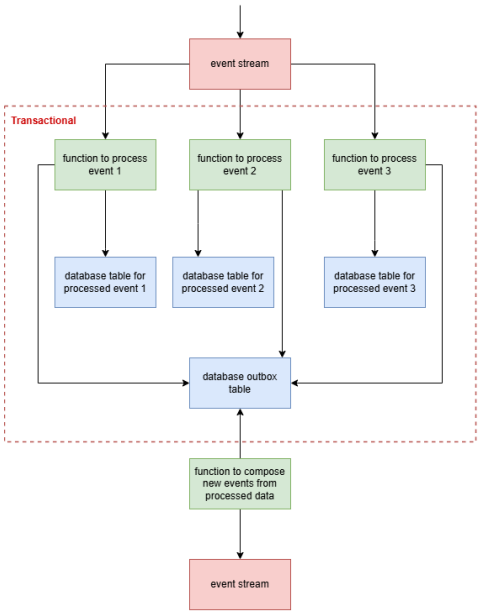


Abb. 2: Konzeptionelle Architektur zur serverlosen Ereignisverarbeitung [1]

Das Konzept ist bewusst technologieoffen gehalten, damit verschiedene Umsetzungsansätze, wie etwa die Umsetzung mit oder ohne Framework, gegenübergestellt und anhand von Kriterien wie Startzeit, Ressourcenverbrauch und Entwicklungsaufwand evaluiert werden können.
Abschließend sollen geeignete Messverfahren integriert werden, um einen präzisen Vergleich zwischen der containerisierten und der serverlosen Variante zu ermöglichen. So lässt sich das Potenzial serverloser Ausführungsmodelle systematisch bewerten.

Umsetzung

Für die Umsetzung sind bereits konkrete Zieltechnologien vorgegeben. Als Function-as-a-Service-Plattform soll AWS Lambda verwendet werden. Die Datenpersistenz erfolgt weiterhin über eine PostgreSQL-

Datenbank, die über Amazon RDS bereitgestellt wird. Für die Eventverarbeitung kommt wie in der Ausgangsanwendung Apache Kafka zum Einsatz.
Die Entwicklung erfolgt mithilfe des Frameworks AWS Serverless Application Model (SAM), das insbesondere die lokale Entwicklung und das Testen von Funktionen unterstützt. Zur Ressourcenschonung und zur Beschleunigung von Entwicklungszyklen werden erste Funktionen lokal ausgeführt. Die vollständige Evaluation der Zielarchitektur ist zu einem späteren Zeitpunkt im produktiven Cloud-Deployment vorgesehen. Durch den Einsatz von AWS SAM wird auch dieser Schritt unterstützt.

Ausblick

Im weiteren Verlauf der Arbeit steht die schrittweise Umsetzung der Cloud-Funktionen im Vordergrund. Ziel ist es, die vollständige Ereignisverarbeitung auf Basis von AWS Lambda funktionsfähig nachzubauen und mit der bestehenden containerbasierten Lösung vergleichbar zu machen. Darüber hinaus sollen verschiedene Umsetzungsvarianten theoretisch evaluiert und dokumentiert werden, um deren Auswirkungen auf Entwicklungsaufwand, Wartbarkeit und Ressourcenverbrauch analysieren zu können.
Begleitend zur technischen Umsetzung ist die Durchführung der geplanten Messungen vorgesehen, um eine objektive Bewertung beider Ausführungsmodelle zu ermöglichen. Die daraus gewonnenen Erkenntnisse sollen eine Einschätzung der Eignung serverloser Architekturen in der vorliegenden Anwendung erlauben und als Grundlage für Empfehlungen hinsichtlich zukünftiger Migrationen und Projekte dienen.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Spiceworks Inc. Adoption of information technology (IT) trends, current and planned, in organizations in North America and Europe for 2025. https://www.statista.com/statistics/1420303/north-america-europe-tech-trend-adoption-2024/, 11 2024.

[3] Samuel Kounev, Nikolas Herbst, Cristina L. Abad, Alexandru Iosup, Ian Foster, Prashant Shenoy, Omer Rana, and Andrew A. Chien. Serverless Computing: What It Is, and What It Is Not? *Communications of the ACM*, 66:80–92, 2023.

[4] Johannes Manner. A Structured Literature Review Approach to Define Serverless Computing and Function as a Service. In *2023 IEEE 16th International Conference on Cloud Computing (CLOUD)*, volume 16, pages 516–522. IEEE, 2023.

[5] Cogent Infotech Pvt Ltd. The Evolution of Cloud Computing: Trends and Emerging Technologies Shaping 2025. https://www.cogentinfo.com/resources/the-evolution-of-cloud-computing-trends-and-emerging-technologies-shaping-2025, 02 2025.

[6] Chris Richardson. *Microservices Patterns: With Examples in Java*. Manning Publications Co., 2019.

[7] Marcia Villalba. Lifting and shifting a web application to AWS Serverless: Part 1. https://aws.amazon.com/de/blogs/compute/lifting-and-shifting-a-web-application-to-aws-serverless-part-1/, 09 2022.

Monokulare Tiefenschätzung zur Entfernungsmessung im Beachvolleyball

Edwin Starz **Steffen Schober**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einführung

Tiefenschätzung aus einfachen RGB-Bildern – ohne aufwendige Sensorik – ist ein relevantes Anwendungsfeld aktueller KI-Forschung. In diesem Projekt wird ein Verfahren entwickelt, das die Entfernungen von Spieler:innen und Ball in Beachvolleyball-Videos automatisch aus Einzelbildern schätzt und für Analysezwecke annotiert.

Zielsetzung der Arbeit

Das Ziel der Arbeit ist, durch die Kombination aus Objekterkennung und Tiefenschätzung die Spielanalyse automatisch um metrische Informationen zu erweitern. Neben Einzelbildverfahren wird auch die Validierung durch LiDAR-Daten durchgeführt, um die Genauigkeit zu prüfen und das Verfahren zu verbessern.

Technologischer Hintergrund

Monokulare Tiefenschätzung ist die Bestimmung der relativen oder absoluten Tiefe pro Pixel aus einem einzelnen RGB-Bild. Klassische Methoden basieren auf Multi-View-Geometrie [7], moderne Modelle hingegen verwenden neuronale Netze, um Tiefenkarten direkt zu schätzen.
Wie in der monokularen Tiefenschätzung üblich, werden folgende Fehlermaße verwendet (niedrigere Werte sind besser):

- **AbsRel** (Absolute Relative Error): Mittlere absolute Relativabweichung
- **RMSE** (Root Mean Square Error): Quadratwurzel des mittleren quadrierten Abstands.
- **MAE** (Mean Absolute Error): Mittlerer absoluter Abstandsfehler, typischerweise in Metern [5].

Zu den gängigen Benchmarks zählen KITTI und NYUv2 [8].
In diesem Projekt kommen folgende Komponenten zum Einsatz:

- **Objekterkennung:** YOLO in einer aktuellen Version, z.B. YOLOv11s [6] für die Erkennung von Spieler:innen und Ball.
- **Tiefenschätzung:** Verschiedene Modelle im Vergleich. Eine Option ist Depth Pro, ein „Foundation Model“, das hochauflösende Tiefenkarten mit korrektem Maßstab liefert [2].
- **Fusion/Validierung:** LiDAR-Punktwolken zur Evaluierung der Schätzungen.
- **Containerisierung:** Docker-Umgebungen [4] zur Reproduzierbarkeit.

Setup und Methodik

Die Kameraaufnahmen stammen aus handelsüblichen Kameras, z.B. Smartphones. Zusätzlich werden Punktwolken einer Hesai Pandar64 LiDAR Kamera aufgenommen. Die Schritte im Überblick:

1. **Bildanalyse:** Mit YOLO [6] werden die Spieler:innen und der Ball lokalisiert.
2. **Tiefenschätzung:** KI-Modelle liefern eine Tiefenkarte zum jeweiligen Frame.
3. **Entfernungsermittlung:** Anhand der Objekterkennung wird der Tiefenwert an den entsprechenden Koordinaten extrahiert.
4. **Annotation:** Mittels eines Python-Skripts wird der Tiefenwert in Metern an dem zugehörigen Objekt angezeigt. 1
5. **Validierung:** LiDAR-Aufnahmen in Form von 3D-Punktwolken dienen als Ground Truth.
6. **Visualisierung:** Die Abweichungen zwischen geschätzter und realer Tiefe (z.B. durch Heatmaps). Bei signifikanten Fehlern kann das Fine-tuning oder die Wahl verschiedener Referenzpunkte in Betracht gezogen werden.

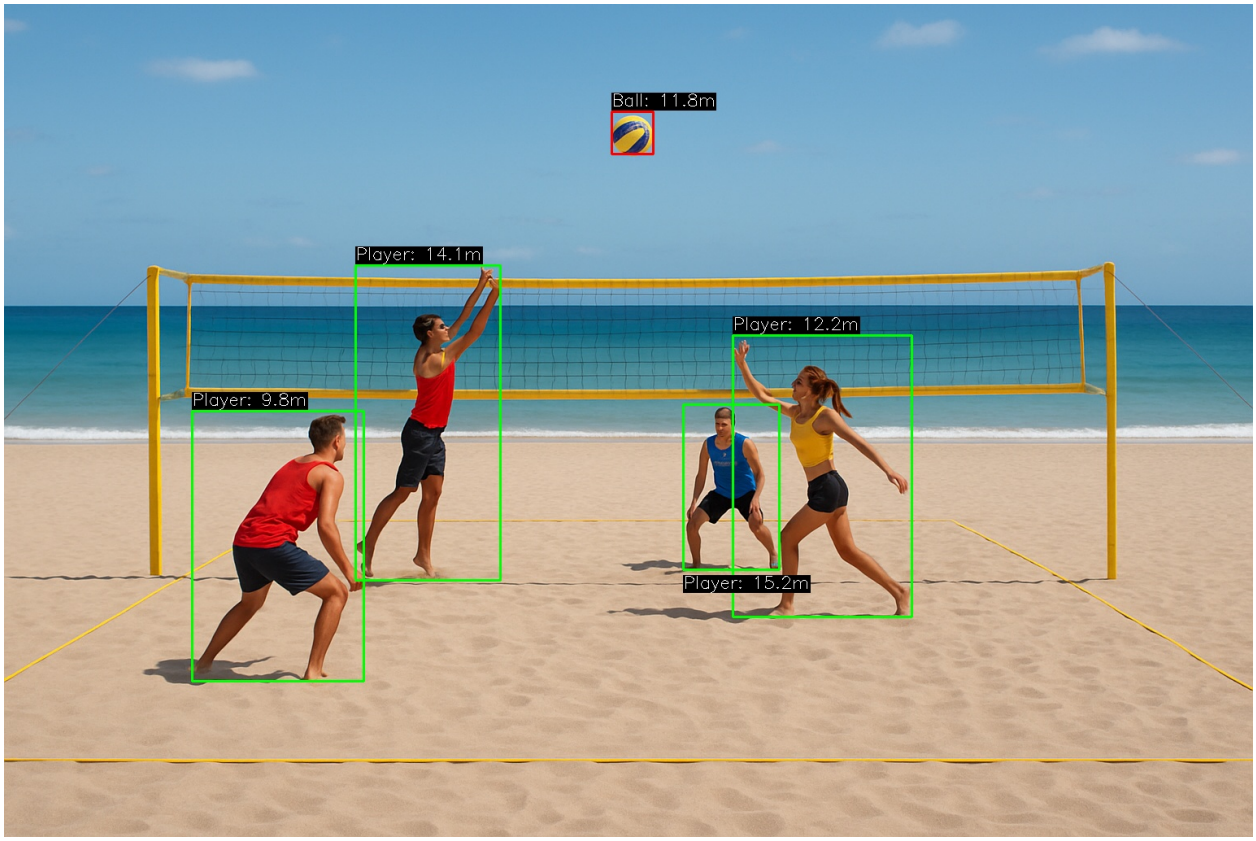


Abb. 1: Beachvolleyball Szene mit Tiefenschätzwerten [3]

Technische Herausforderungen

Eine zentrale Herausforderung stellt die Datenfusion dar: Die intrinsischen Parameter der Kamera müssen bekannt sein, um beispielsweise die Verzerrung der Linse zu berücksichtigen. Dies wird üblicherweise mit Aufnahmen eines Schachbrettmusters mit bekannten Kantenlängen durchgeführt. Kamerabilder und LiDAR-Punktwolken müssen präzise synchronisiert und extrinsisch kalibriert werden. Hierfür wird das Open-Source-Framework velo2cam_calibration [1] genutzt, das, typischerweise in ROS-Umgebungen betrieben, die exakte räumliche Beziehung zwischen den Sensoren bestimmt. Damit lassen sich das Kamerabild und die Punktwolke übereinanderlegen 2.
Die Auswahl des geeigneten Tiefenschätzmodells ist ebenfalls kritisch. Neben der reinen Genauigkeit auf Testdatensätzen ist die zeitliche Konsistenz der Schätzungen über Videosequenzen hinweg entscheidend, um visuelles Flackern in der Distanzanzeige zu vermeiden. Containerisierung mit Docker stellt sicher, dass die Entwicklungsumgebung inklusive aller Software-Abhängigkeiten exakt dokumentiert und portierbar ist, was ein entscheidender Faktor für die Reproduzierbarkeit und spätere Weiterverwendung ist.

Ausblick

Erweiterung auf die Betrachtung vorheriger und nachfolgender Frames, um Bewegungen besser und konsistenter zu erfassen. Auch eine automatische Spielzugerkennung ist interessant für die Spielanalyse. Weitere Schritte beinhalten:

- Verbesserung durch Spielfeld-/Netzsegmentierung
- Projektion in eine Top-Down-Perspektive

Langfristig könnte ein „End-to-End“-Modell entstehen, das die Tiefenschätzung, Objekterkennung und Distanzbestimmung in einem gemeinsamen Netzwerk erledigt – validiert durch echte Sensordaten.

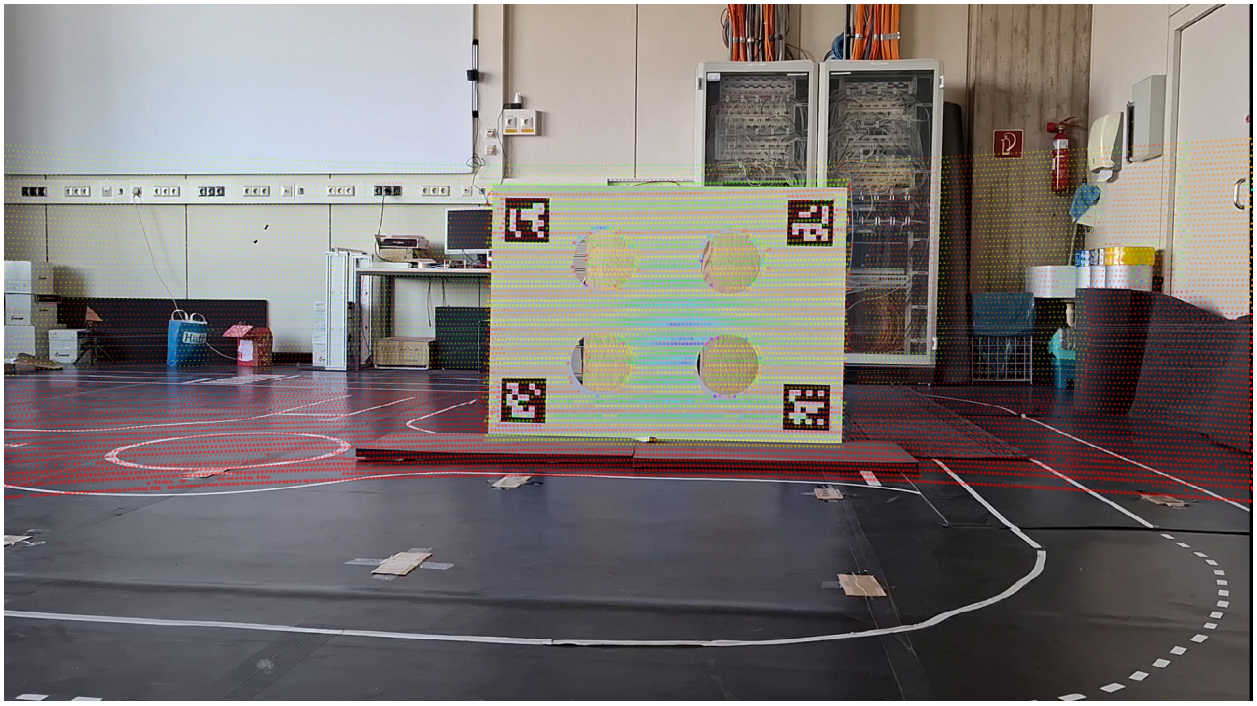


Abb. 2: RGB-Bild und LiDAR Punktewolke fusioniert [3]

Literatur und Abbildungen

[1] Jorge Beltrán, Carlos Guindel, Arturo de la Escalera, and Fernando García. Automatic Extrinsic Calibration Method for LiDAR and Camera Sensor Setups. https://github.com/beltransen/velo2cam_calibration/tree/master, 2022.

[2] Aleksei Bochkovskii, Amael Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan Richter, and Vladlen Koltun. Depth Pro: Sharp Monocular Metric Depth in Less Than a Second. https://arxiv.org/abs/2410.02073, 2025.

[3] Eigene Darstellung.

[4] Docker Inc. Docker Documentation – What is Docker? https://docs.docker.com/get-started/overview/, 2024.

[5] Faisal Khan, Saqib Salahuddin, and Hossein Javidnia. Deep Learning-Based Monocular Depth Estimation Methods—A State-of-the-Art Review. https://www.mdpi.com/1424-8220/20/8/2272, 2020.

[6] Muhammad Hussain Rahima Khanam. YOLOv11: An Overview of the Key Architectural Enhancements. *arXiv*, 2024.

[7] Ashutosh Saxena, Min Sun, and Andrew Y. Ng. Make3D: Learning 3D Scene Structure from a Single Still Image. *IEEE*, pages 824–840, 2008.

[8] Igor Vasiljevic et al. DIODE: A Dense Indoor and Outdoor DEpth Dataset. https://arxiv.org/abs/1908.00463, 2019.

Power BI Implementierung auf Grundlage der Praxisanforderungen eines kleinen- und mittelständischen Unternehmens

Oezguer Uenlue **Dirk Hesse**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Fripac-Medis GmbH, Spiegelberg

Einleitung

Mit der zunehmenden Digitalisierung wächst auch die Bedeutung, Unternehmensdaten gezielt auszuwerten und für Entscheidungen nutzbar zu machen. Business Intelligence (BI) bietet hierfür hilfreiche Werkzeuge, stellt aber besonders kleine und mittelständische Unternehmen (KMU) vor organisatorische und technische Herausforderungen.
Ein Beispiel ist die Fripac-Medis GmbH aus Baden-Württemberg, die professionellen Friseurbedarf vertreibt. Das Unternehmen bietet eine breite Produktpalette für Kunden im In- und Ausland.
Diese Arbeit untersucht die Einführung von Power BI bei Fripac-Medis. Im Mittelpunkt stehen Analysen zu Umsatz, Deckungsbeitrag, Preis, Absatz und offenen Aufträgen. Ziel ist es zu zeigen, wie Power BI als benutzerfreundliches BI-Tool zur datenbasierten Entscheidungsunterstützung in einem KMU genutzt werden kann.

Business Intelligence

Unternehmen nutzen Datenbanken einerseits zur Erfassung alltäglicher Geschäftsprozesse, etwa zur Dokumentation von Zahlungen oder zur Nachverfolgung von Aufträgen. Andererseits können die gesammelten Daten gezielt ausgewertet werden, um wichtige Erkenntnisse für unternehmerische Entscheidungen zu gewinnen.
Ein Beispiel dafür ist ein Restaurantbetrieb in Los Angeles: Durch die Analyse von Kreditkartendaten wurde erkannt, dass viele Gäste gut ausgebildet waren, Wert auf Qualität legten und gerne guten Wein tranken. In der Folge wurde das Angebot gezielt angepasst – etwa durch mehr vegetarische Gerichte und hochwertige Weine – was zu einer deutlichen Umsatzsteigerung führte.
Während solche Analysen in kleineren Unternehmen relativ einfach umsetzbar sind, benötigen größere Unternehmen mit komplexeren Datenstrukturen spezialisierte Systeme wie Data Warehouses oder Data Marts, die große Datenmengen aus verschiedenen Quellen konsolidieren. Ergänzend helfen Techniken wie Data Mining und mehrdimensionale Datenanalysen dabei, Muster und Zusammenhänge zu erkennen. Diese Methoden werden unter dem Begriff Business Intelligence (BI) zusammengefasst und unterstützen datenbasierte Entscheidungen. [1]

Abb. 1: Business-Inteligence [2]

Power BI

Power BI ist eine Plattform von Microsoft, die Softwaredienste, Apps und Konnektoren kombiniert, um Daten aus unterschiedlichen Quellen in interaktive Visualisierungen zu überführen. Die Daten können aus einfachen Excel-Tabellen, cloudbasierten oder lokalen Systemen stammen. Power BI ermöglicht es, Daten zu verknüpfen, interaktive Berichte zu erstellen und diese mit anderen zu teilen.
Die Plattform besteht aus drei zentralen Komponenten: 1. **Power BI Desktop** – Eine Windows-Anwendung zur Erstellung und Analyse von Berichten. 2. **Power BI-Dienst** – Ein Online-Service zum Teilen und Veröffentlichen von Dashboards. 3. **Mobile Power BI-**

Apps – Anwendungen für Smartphones und Tablets zur mobilen Nutzung.
Dank dieser Struktur eignet sich Power BI besonders für KMUs, die kostengünstig und flexibel Business-Intelligence-Funktionen nutzen möchten. [3]

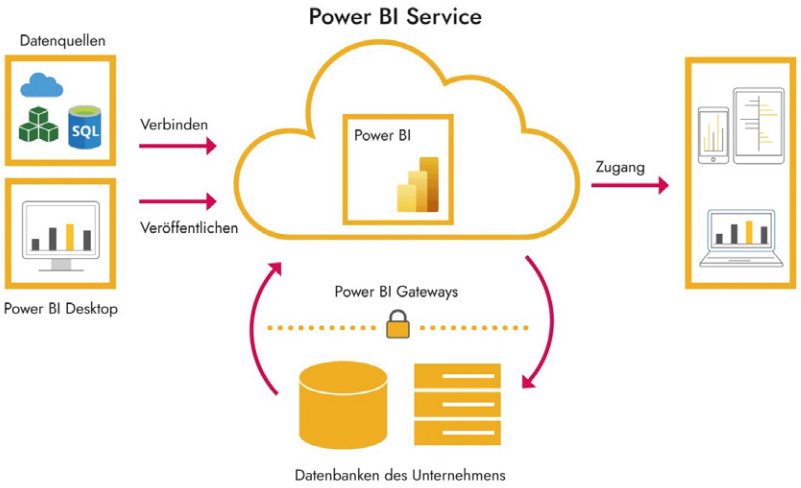


Abb. 2: Power BI-Hauptkomponenten [3]

Anforderungen

Umsatzanalyse

Der Fokus liegt zunächst auf der Umsatzanalyse. Visualisiert wird der aktuelle Gesamtumsatz, ergänzt durch eine Aufschlüsselung nach Abteilungen. Um zeitliche Entwicklungen nachvollziehen zu können, werden die Umsätze zusätzlich auf monatlicher, quartalsweiser und täglicher Basis dargestellt. Darüber hinaus erfolgt eine Auswertung nach Debitoren – also nach einzelnen Kunden oder Kundengruppen – sowie eine differenzierte Darstellung nach Marken und Verbänden. Diese Perspektiven ermöglichen es, umsatzstarke Segmente gezielt zu identifizieren.

Deckungsbeitragsanalyse

Anschließend folgt die Darstellung des Deckungsbeitrags. Die Daten lassen sich nach Marken und Kategorien filtern. Grundlage der Analyse sind die Einkaufspreise und weiteren Kosten, aus denen der Deckungsbeitrag abgeleitet wird. Diese Kennzahl liefert Informationen über die Rentabilität einzelner Produkte oder Gruppen. Die visuelle Darstellung unterstützt dabei, wirtschaftlich erfolgreiche Produkte zu erkennen und weniger rentable Bereiche zu optimieren.

Absatz- und Preisanalyse

Im weiteren Verlauf widmet sich das Dashboard der Absatz- und Preisanalyse. Auch hier ist eine Filterung nach Marken und Kategorien möglich. Gezeigt werden die Verkaufspreise aus dem Vorjahr und dem aktuellen Jahr, ergänzt durch die prozentuale Preisentwicklung. Zusammen mit den Absatzzahlen lassen sich Aussagen zur Preisentwicklung und zur Preiselastizität einzelner Produkte ableiten. Diese Informationen sind wichtig für eine fundierte Preisstrategie.

Auftragsübersicht

Abschließend wird eine Übersicht über sämtliche Kundenaufträge präsentiert. Die Daten lassen sich nach Datum oder Käufer filtern. Angezeigt werden jeweils die gekauften Produkte, die Menge, der Einzelpreis sowie der Endpreis des jeweiligen Auftrags. Diese strukturierte Darstellung erleichtert die Analyse von Verkaufsaktivitäten und ermöglicht eine schnelle Übersicht über umsatzstarke Kunden, Produkte oder Zeiträume.

Fazit

Die Implementierung von Power BI zur Analyse und Visualisierung von Unternehmensdaten bietet kleinen und mittelständischen Unternehmen eine praxisnahe und leistungsstarke Lösung. Durch die systematische Aufbereitung von Kennzahlen wie Umsatz, Deckungsbeitrag, Preisentwicklung und Aufträgen erhalten Unternehmen fundierte Entscheidungsgrundlagen. Power BI ermöglicht es, große Datenmengen übersichtlich darzustellen und flexibel auf unterschiedliche Analysebedürfnisse einzugehen. Insgesamt zeigt sich, dass Power BI für KMUs eine benutzerfreundliche und effektive Plattform darstellt, um datenbasierte Entscheidungen zu treffen und die Effizienz der Unternehmenssteuerung zu steigern.

Literatur und Abbildungen

[1] Kenneth C. Laudon, Jane Price Laudon, and Detlef Schoder. *Wirtschaftsinformatik eine Einführung*. Pearson Studium, 3 edition, 2016.

[2] Raghavan Srinivas. Business Intelligence and Analytics – Part I of II. https://www.gauri.com/business-intelligence-and-analytics-part-i-of-ii/, 03 2024.

[3] Laurenz Wuttke. Was ist Power BI? Leitfaden für Einsteiger in PowerBI. https://datasolut.com/was-ist-power-bi/, 08 2024.

Entwicklung einer Best-Practice-Richtlinie für den Einsatz generativer KI beim Erlernen einer Programmiersprache

Raluca Maria Vedislav **Dennis Grewe**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Motivation

Die zunehmende Verbreitung generativer Künstlicher Intelligenz (KI) verändert den Lernprozess im Hochschulbereich grundlegend und stellt traditionelle Lehrmethoden vor neue Herausforderungen [2]. Insbesondere beim Erlernen einer neuen Programmiersprache eröffnen KI-gestützte Tools eine Vielzahl neuer Möglichkeiten, um Studierende in ihrem Lernprozess zu unterstützen. Intelligente Codevervollständigungsdienste wie GitHub Copilot helfen dabei, Syntaxfehler zu vermeiden und effizienteren Code zu schreiben, indem sie kontextbezogene Vorschläge liefern. Chatbots und Large Language Models (LLMs) wie Open AI ChatGPT oder Google Gemini können Fragen beantworten, Erklärungen zu Konzepten liefern oder Code analysieren und verbessern [4].

Darüber hinaus ermöglichen KI-basierte Tutoren eine adaptive Lernumgebung, die sich individuell an den Fortschritt und die Bedürfnisse der Studierenden anpasst. Beispielsweise können personalisierte Übungsaufgaben generiert oder detaillierte Schritt-für-Schritt-Anleitungen für komplexe Programmierprobleme gegeben werden [5]. Auch Debugging-Prozesse werden durch KI vereinfacht, indem Fehler automatisiert erkannt, mögliche Lösungen vorgeschlagen und alternative Implementierungen aufgezeigt werden [4].

Trotz dieser Vorteile besteht jedoch die Gefahr, dass Studierende, die sich vor allem in den Anfängen ihrer Programmierkarriere befinden, wie beispielsweise im Grundstudium, sich zu stark auf KI verlassen und die grundlegenden Problemlösungsfähigkeiten sowie das algorithmische Denken vernachlässigen. Eine unreflektierte Nutzung kann dazu führen, dass Code lediglich kopiert wird, ohne das zugrunde liegende Konzept vollständig zu verstehen. Dies kann langfristig negative Auswirkungen auf das eigenständige Denken und die Fähigkeit zur kreativen Lösung von Programmierproblemen haben [3]. Daher ist es entscheidend, Methoden zu entwickeln, die eine ausgewogene Nutzung von KI ermöglichen und sicherstellen, dass Studierende die Technologie als unterstützendes Werkzeug und nicht als Ersatz für das eigene Denken betrachten.

Zielsetzung

Diese Arbeit verfolgt das Ziel, eine Best-Practice-Richtlinie für den gezielten Einsatz von KI beim Erlernen einer neuen Programmiersprache zu entwickeln und deren Wirksamkeit empirisch zu untersuchen. Die Richtlinie richtet sich insbesondere an Studierende im Grundstudium, die sich im ersten Semester mit den Grundlagen des Programmierens auseinandersetzen. Sie ermöglicht eine reflektierte Nutzung von KI beim Programmierenlernen und gibt konkrete Empfehlungen dazu, wann und wie KI sinnvoll eingesetzt werden kann, um den Lernprozess zu unterstützen, ohne das eigenständige Denken zu untergraben.

Methodik

Zur Entwicklung der Richtlinie wird zunächst ein fundierter theoretischer Hintergrund erarbeitet. Dieser umfasst die Rolle von Künstlicher Intelligenz im Bildungskontext, ihre Potenziale für personalisiertes Lernen sowie die damit verbundenen Herausforderungen. Zudem werden typische Schwierigkeiten beim Erlernen von Programmiersprachen analysiert und bestehende didaktische Konzepte zur Unterstützung des Lernprozesses betrachtet. Ein weiterer Fokus liegt auf der Untersuchung KI-gestützter Lernmethoden wie Chatbots, Codevervollständigungsdiensten oder intelligenten Tutoren, die beim Programmierenlernen zum Einsatz kommen können. Ergänzend werden bestehende Ansätze zur strukturierten und reflektierten Nutzung von KI im Bildungsbereich analysiert, um darauf aufbauend eine eigene Richtlinie zu entwickeln. Zur Evaluierung der Richtlinie wird eine empirische Untersuchung im Kurs „Programmieren“ in den informatikverwandten Studiengängen an der Hochschule Esslingen durchgeführt. Die Untersuchung folgt einem experimentellen Design mit zwei Gruppen:

- **Gruppe A (Kontrollgruppe):** Nutzt KI ohne Einschränkungen und ohne zusätzliche Anleitung für effiziente Interaktion.
- **Gruppe B (Experimentalgruppe):** Erhält die Best-Practice-Richtlinie als Unterstützung sowie eine kurze Einweisung zur strukturierten KI-Nutzung.

Beide Gruppen bearbeiten die gleiche Programmieraufgabe mit KI-Unterstützung. Die Untersuchung umfasst drei Phasen:

- **Pre-Test:** Vor der Aufgabenbearbeitung wird ein Fragebogen ausgefüllt, um die Vorerfahrungen der Studierenden, ihre bisherige Nutzung von KI sowie ihr Vorwissen anhand einiger Kompetenzfragen zu bereits beareiteteten Themen aus der Vorlesung zu erfassen.
- **Aufgabenbearbeitung:** Die Studierenden erhalten 90 Minuten Zeit, um eine Programmieraufgabe umzusetzen. Dabei handelt es sich um das Spiel „Snake“, das als Konsolenspiel in C programmiert werden soll. Die Programmieraufgabe ist dem Wissenstand der Studierenden zum aktuellen Zeitpunkt angepasst. Der Lösungsraum ist offen, sodass Studierende Ihr Wissen aber auch Vorschläge der KI beim Lösen der Problemstellung einbringen können.
- **Post-Test:** Nach der Aufgabenbearbeitung beantworten die Teilnehmenden einen weiteren Fragebogen, um die subjektiven Erfahrungen mit der KI-Nutzung, das Verständnis der gelösten Aufgabe und die wahrgenommene Unterstützung durch die Best-Practice-Richtlinie zu erheben.

Die Bearbeitung erfolgt in einem vorgegebenen Setup mit Visual Studio Code als IDE und bwGPT als unterstützendem KI-Tool. Da Visual Studio Code bereits in der Lehrveranstaltung verwendet wird, arbeiten die Studierenden in einer vertrauten Entwicklungsumgebung und müssen sich nicht auf neue Werkzeuge einstellen. bwGPT, das auf GPT-4o basiert, wird von der Hochschule Esslingen bereitgestellt und verarbeitet ausschließlich technisch notwendige Daten. Die Kommunikation mit dem Tool erfolgt über sogenannte Prompts, also natürliche Spracheingaben, mit denen Nutzer gezielt Fragen stellen, Aufgaben beschreiben oder Code generieren lassen können. Aufgrund der datenschutzfreundlichen Ausrichtung fiel die Wahl auf dieses Tool.
Abbildung 1 zeigt eine Übersicht der Methodik, die angewendet wurde, mit der Aufteilung der Studierenden in zwei Gruppen, dem Pre-Test vor der Programmieraufgabe sowie dem Post-Test danach.

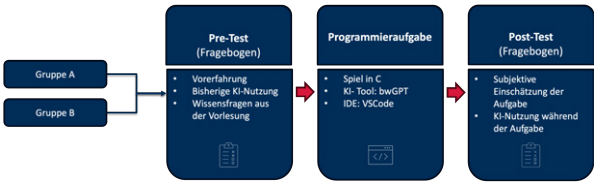


Abb. 1: Darstellung der Methodik [1]

Zusätzlich geben die Teilnehmenden den erstellten Code sowie die verwendeten Prompts ab, um die Qualität der Lösungen und den Umgang mit KI genauer analysieren zu können.

Datenanalyse

Nach Abschluss der Erhebung erfolgt eine Auswertung der gesammelten Daten. Die Analyse konzentriert sich auf qualitative Verfahren, es werden aber auch quantitative Verfahren durchgeführt.
Die qualitative Analyse konzentriert sich auf die Auswertung der abgegebenen Programmcodes und Prompts. Dabei werden sogenannte *Code Smells* sowie *Code Metrics* herangezogen, um die Qualität und Struktur der Lösungen zu bewerten. Zusätzlich wird untersucht, wie die KI eingesetzt wurde – insbesondere im Hinblick auf die Formulierung der Texteingaben und der Konversation mit der KI und die Anwendung der Empfehlungen aus der Best-Practice-Richtlinie. Zur systematischen Auswertung werden Kategorien definiert und die Ergebnisse im Rahmen einer Inhaltsanalyse interpretiert.
Die quantitativen Daten aus den Pre- und Post-Tests (z. B. Selbsteinschätzung, Nutzungshäufigkeit von KI, Verständnis, wahrgenommene Unterstützung) werden mittels statistischer Verfahren analysiert. Dabei werden Kennzahlen ermittelt, um Unterschiede zwischen der Kontroll- und der Experimentalgruppe zu identifizieren. Ziel ist es, mögliche Effekte der Best-Practice-Richtlinie auf den Lernerfolg und die wahrgenommene Unterstützung durch die KI zu überprüfen.
Die Kombination beider Analyseformen ermöglicht eine Bewertung der Wirksamkeit der entwickelten Richtlinie. Sie liefert Erkenntnisse darüber, inwiefern die strukturierte KI-Nutzung das Verständnis fördert, zur besseren Aufgabenlösung beiträgt und wie bewusst die KI eingesetzt wurde.

Richtlinie

Die entwickelte Best-Practice-Richtlinie dient als praxisorientierter Leitfaden für den reflektierten Einsatz von KI beim Programmierenlernen. Sie richtet sich an Einsteigende und ist so aufgebaut, dass sie leicht verständlich und direkt anwendbar ist. Zu Beginn bietet die Richtlinie eine Einführung in das Thema

KI-gestütztes Lernen. Dabei werden die grundlegenden Begriffe erklärt, der Nutzen und die Grenzen von KI-Tools im Lernkontext aufgezeigt und der typische Lernverlauf im Umgang mit KI beim Programmieren beschrieben. Ergänzt wird dies durch eine Übersicht über verschiedene Arten von KI-Modellen, ihre Einsatzgebiete und einen Fokus auf Large Language Models (LLMs), wie z. B. GPT, Claude oder GitHub Copilot.
Im Hauptteil folgen konkrete Hinweise und Handlungsempfehlungen für typische Situationen beim Programmierenlernen, in denen KI sinnvoll eingesetzt werden kann. Besonderes Augenmerk gilt hierbei auch den Fallstricken im Umgang mit KI, wie etwa der Halluzinationsgefahr, mangelnder Reproduzierbarkeit von Ergebnissen oder der Notwendigkeit zur Übernahme von Verantwortung durch die Lernenden.
Ein zentraler Bestandteil der Richtlinie ist das Kapitel Prompt Engineering. Es zeigt, wie durch gezielte und gut strukturierte Prompts qualitativ hochwertige und relevante KI-Antworten erzielt werden können. Neben einer Einführung in die Bestandteile eines Prompts werden verschiedene Prompting Verfahren vorgestellt. Darüber hinaus wird auf typische Herausforderungen beim Prompting eingegangen.

Mithilfe dieser Richtlinie wird Studierenden ein strukturierter und reflektierter Einstieg in die Nutzung von KI-Tools beim Programmierenlernen ermöglicht. Sie fördert ein grundlegendes Verständnis für die Funktionsweise von KI sowie die Fähigkeit, diese gezielt und verantwortungsvoll im Lernprozess einzusetzen.

Ausblick

Langfristig soll die entwickelte Richtlinie nicht nur im Rahmen der vorliegenden Untersuchung Anwendung finden, sondern auch als Grundlage für weiterführende didaktische Konzepte im Programmierenlernen mit KI dienen. Denkbar ist eine Integration in Lehrveranstaltungen oder Tutorien, um Studierende frühzeitig für einen reflektierten Umgang mit KI-Tools zu sensibilisieren. Zukünftige Arbeiten könnten die Richtlinie auf andere Programmiersprachen oder Anwendungsfelder ausweiten und die Auswirkungen auf verschiedene Lerntypen noch differenzierter untersuchen. Auch die kontinuierliche Anpassung an neue technologische Entwicklungen im Bereich der KI bleibt dabei ein zentrales Ziel.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Mohammad Amin Kuhail, Nazik Alturki, Salwa Alramlawi, et al. Interacting with educational chatbots: A systematic review. *Education and Information Technologies*, 28:973–1018, 2023.

[3] Manuel Rene Theise. ChatGPT: Risiken, Gefahren und Chancen in Lehre und Forschung. *WiST Zeitschrift für Studium*, pages 17–23, 2023.

[4] Pedro Wightman. Twisted Games: A First Experience of Inclusion of AI tools in First Year Programming Classes. In *2024 IEEE Latin American Conference on Computational Intelligence, LA-CCI 2024 - Proceedings*. Institute of Electrical and Electronics Engineers Inc., 2024.

[5] Sualeha Zafar, Farzana Shaheen, and Javaria Rehan. Use of ChatGPT and Generative AI in Higher Education: Opportunities, Obstacles and Impact on Student Performance. *iRASD Journal of Educational Research*, 5:01–12, 2024.

Datengetriebenes Unternehmen: Integration eines BI-gestützten Beratungsansatzes in die Fahrzeugproduktionsprozesse der Mercedes-Benz AG

Walter Vins **Anke Bez**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Mercedes-Benz AG, Sindelfingen

Einleitung

Aufgrund von Industrialisierung, Globalisierung, Umweltauflagen und hohen Sicherheitsanforderungen herrscht in der Automobilindustrie ein intensiver Wettbewerb [1]. Unternehmen sind gefordert, ihre Produktionsprozesse flexibel zu gestalten und fundierte Entscheidungen auf Basis aktueller Daten zu treffen. Die Inhouse Lean Beratung der Mercedes-Benz AG verfolgt deshalb einen datenbasierten Ansatz, bei dem Business Intelligence gezielt eingesetzt wird, um klassische Lean-Prinzipien wie die Vermeidung von Verschwendung und die kontinuierliche Verbesserung mithilfe datengetriebener Analysen gezielt weiterzuentwickeln und die Effektivität der Beratung zu erhöhen.

Problemstellung und Zielsetzung

Die Umsetzung datenbasierter Arbeitsweisen stellt viele Unternehmen vor erhebliche Herausforderungen. Mangelhafte Datenqualität, fehlendes analytisches Fachwissen, Bedenken hinsichtlich Data Governance und Sicherheit, eingeschränkter Zugang zu relevanten Daten sowie fehlende Unterstützung durch das Top-Management zählen zu den zentralen Hürden auf dem Weg zu einer datengetriebenen Organisation [4].
Ziel dieser Arbeit ist es daher, praxisnahe Strategien zur erfolgreichen Implementierung eines datenbasierten Lean-Beratungskonzepts zu entwickeln. Dieses Ziel wird mit Hilfe eines integrativen Change Management-Prozesses verfolgt, an dessen Ende ein klar definiertes Rollenverständnis sowie konkrete Maßnahmen zur langfristigen Verankerung datengetriebener Arbeitsweisen im Beratungsumfeld steht.

Theoretische Grundlagen

Inhouse Lean Beratung

Lean Management unterstützt die Senkung betrieblicher Kosten und fördert einen ressourcenschonenden Einsatz vorhandener Mittel. Gleichzeitig können sowohl die Leistungsqualität des Unternehmens als auch die Kundenzufriedenheit verbessert werden. Im Mittelpunkt steht die konsequente Ausrichtung unternehmerischer Abläufe auf Tätigkeiten, die aus Sicht der Kunden einen erkennbaren Nutzen stiften [3].
Die Inhouse Lean Beratung der Mercedes-Benz AG verbindet diesen Ansatz mit der Stärke interner Beratungsteams. Durch ihre organisatorische Nähe und das tiefe Verständnis der unternehmensspezifischen Abläufe können sie insbesondere Veränderungsprozesse und Projekte im Bereich der Fahrzeugproduktion wirkungsvoll unterstützen und vorantreiben.

Business Intelligence (BI) und datengetriebene Unternehmen

Datengetriebene Unternehmen verstehen Daten als strategischen Erfolgsfaktor, nutzen sie organisationsweit und systematisch, greifen dabei auf aktuelle oder Echtzeitdaten zurück und treffen auf dieser Basis Entscheidungen [5].
Unterstützen lässt sich dieser Ansatz durch BI, das in klassischer Definition sämtliche informationstechnischen Werkzeuge, Methoden und Verfahren umfasst, die den Umgang mit Daten sowie die Gewinnung und Nutzung von Informationen zur Entscheidungsfindung ermöglichen [7].
Ein entscheidender Erfolgsfaktor ist jedoch die Datenkompetenz der Mitarbeitenden, die sich in der Fähigkeit zeigt, Daten sicher, reflektiert und zielgerichtet zu nutzen, ein Verständnis für datenverarbeitende Prozesse zu entwickeln und daraus gewonnene Informationen sinnvoll in Entscheidungen einzubringen [7].

Je nach Erfahrungsgrad lässt sich die Belegschaft in vier Kompetenzstufen unterteilen. Diese gehen von daten-unbewusst bis daten-spezialisiert, was die gezielte Qualifizierung erleichtert [7].
Wie in Abbildung 1 dargestellt, werden diese Kompetenzstufen in einer Pyramide visualisiert. Sie zeigt die Entwicklung von einem daten-unbewussten Verhalten hin zu einem daten-spezialisierten Umgang und unterstützt so die Planung gezielter Schulungs- und Veränderungsmaßnahmen.

Abb. 1: Datenkompetenz-Entwicklungsstufen in Organisationen (eigene Darstellung in Anlehnung an Weichand) [7]

Change Management

Veränderungen in Unternehmen sind nicht allein technisch zu bewältigen, sondern benötigen ein strukturiertes Change Management. Dieses umfasst die aktive Gestaltung von Übergängen und Veränderungen in Strategie, Prozessen und Kultur [2]. Dabei treten regelmäßig Widerstände auf, die rational, politisch oder emotional begründet sein können und gezielt adressiert werden müssen [6].
Praktische Erfahrungen im Umgang mit Veränderungsprozessen machen deutlich, dass die Reaktionen der betroffenen Mitarbeitenden auf geplante Neuerungen stark variieren. Ein Teil zeigt sich aufgeschlossen und unterstützend, ein weiterer begegnet der Veränderung mit Zurückhaltung oder Unentschlossenheit, während eine dritte Gruppe sich deutlich dagegen positioniert. Diese grundsätzliche Dreiteilung lässt sich durch die genauere Unterscheidung in sieben spezifische Reaktionstypen weiter ausdifferenzieren, wie in Abbildung 2 veranschaulicht wird [6].

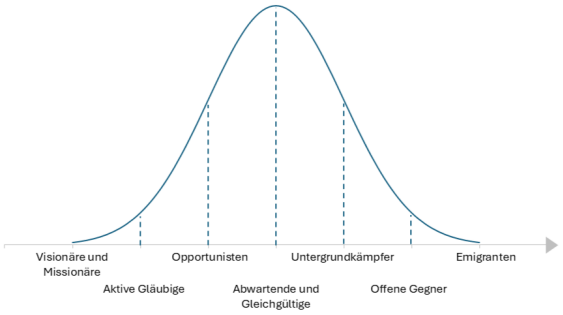


Abb. 2: Reaktionstypen im Veränderungsprozess (eigene Darstellung in Anlehnung an Vahs) [6]

Methodik

Zur fundierten Erarbeitung der Herausforderungen und Lösungsansätze werden qualitative Experteninterviews durchgeführt. Befragt werden Lean-Berater, die als Pioniere in der Anwendung datenbasierter Arbeitsweisen gelten. Ziel ist es, praxisnahe Erkenntnisse über Erfolgsfaktoren, Barrieren und Strategien zur Selbstqualifikation im Bereich BI zu gewinnen. Die Interviews werden halbstrukturiert geführt, um sowohl Vergleichbarkeit als auch die Erfassung individueller Erfahrungen zu ermöglichen.

Ausblick

Die Integration eines BI-gestützten Beratungsansatzes erfordert weit mehr als nur die technische Implementierung eines BI-Tools. Entscheidend sind eine umfassende organisatorische Verankerung, klare Rollenverteilungen, gezielte Qualifizierungsmaßnahmen und die Förderung einer datengetriebenen Unternehmenskultur. In den weiteren Kapiteln der Arbeit werden die Experteninterviews vollständig ausgewertet und vertiefte Handlungsempfehlungen entwickelt. Ziel ist es, ein praxisnahes Modell zur nachhaltigen Verankerung eines BI-gestützten Lean-Beratungsansatzes im Unternehmen zu entwerfen.

Literatur und Abbildungen

[1] Tefi Alonso. How Toyota Went From Humble Beginnings To Automotive Giant. https://www.cascade.app/studies/how-toyota-went-from-humble-beginnings-to-automotive-giant, 2023.

[2] Frank Bertagnolli. *Lean Management: Introduction and In-Depth Study of Japanese Management Philosophy*. Springer Wiesbaden, 1 edition, 2022.

[3] Franz J. Brunner. *Japanische Erfolgskonzepte*. Carl Hanser Verlag, 5 edition, 2023.

[4] Fern Halper and David Stodder. What it takes to be data-driven. *TDWI Best Practices Report*, pages 33–49, 2017.

[5] Jonas Rashedi. *Das datengetriebene Unternehmen: Erfolgreiche Implementierung einer data-driven Organization*. Springer Gabler Wiesbaden, 1 edition, 2022.

[6] Dietmar Vahs. *Organisation: Ein Lehr- und Managementbuch*. Schäffer-Poeschel Verlag, 11 edition, 2023.

[7] Andrea Weichand. *Agile Datenkompetenz: Reporting-Prozesse mit und ohne Excel gestalten*. Springer Gabler Wiesbaden, 1 edition, 2023.

Entwicklung und Implementierung eines Backend-Services zur Verwaltung und Überwachung von E-Scooter Sharing-Flotten

Vincent Vollmer **Mirko Sonntag**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung und Motivation

Im Zuge der Digitalisierung und nachhaltigen Mobilitätslösungen gewinnen vernetzte Fahrzeuge, insbesondere im Bereich der Mikromobilität, zunehmend an Bedeutung. E-Scooter stellen eine flexible und umweltfreundliche Alternative zum Individualverkehr dar. Um deren Betrieb effizient zu gestalten, ist eine zuverlässige, sichere und erweiterbare IT-Infrastruktur notwendig.

Studenten der Hochschule Esslingen nutzen hauptsächlich Nahverkehr oder private Fahrzeuge um zu ihrem Campus zu kommen. Mit dem bevorstehenden Umzug des Standorts Flandernstraße in die Esslinger Weststadt, plant die Hochschule in Zusammenarbeit mit dem Anwendungszentrum KEIM des Fraunhofer Instituts eine eigene E-Scooter-Forschungsflotte für das KEIM anzubieten. Diese dient als flexible, kostengünstige Alternative im Nahverkehr.

Ziel dieser Arbeit ist es, ein Backend-System zu entwickeln, das eine stabile Kommunikation zwischen E-Scootern und einem zentralen Cloud-Server ermöglicht. Zusätzlich sollen Benutzeranfragen über eine Web- oder Mobile-Client-Anwendung verarbeitet werden, wobei der Authentifizierungsmechanismus Keycloak zum Einsatz kommt.

Das Backend dient den Studenten in Kombination mit einem entsprechenden Frontend als mobile Buchungs-App. Mit dieser können E-Scooter für bestimmte Zeiten gebucht und zum Fahren freigeschalten werden.

Die Anbindung der Mikromobilitätsfahrzeuge an den Server ist somit ebenfalls wesentlich. Andauernde Übermittlung von Standortdaten sowie anderen Fahrzeugparametern sind maßgeblich für diese Kommunikation mit einem eingebauten TCP-Server.

Umsetzung

Abbildung 1 gibt einen Überblick über das gesamte System mit allen Softwarekomponenten.

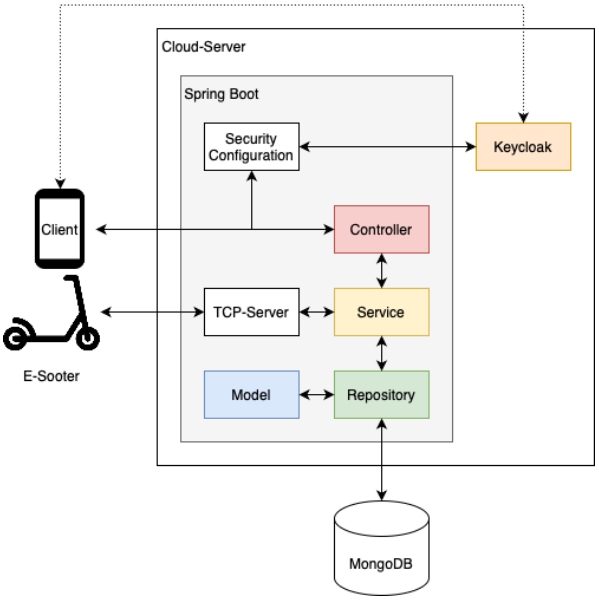


Abb. 1: Komponentendiagramm [1]

Im Zentrum steht der Cloud-Server, welcher über zwei Hauptschnittstellen kommuniziert: einerseits mit den E-Scootern über eine TCP-Verbindung, andererseits mit den Clients (z. B. mobile Apps) über HTTP-Endpunkte.

Die entwickelte Systemarchitektur basiert auf einer Spring Boot Anwendung, die in der Cloud-Umgebung betrieben wird. Die Applikation wird mithilfe des Spring Boot Frameworks implementiert und beinhaltet eine REST API (Representational State Transfer Application Programming Interface). Mithilfe dieser Schnittstelle können Webdienste im Internet einfach und schnell erstellt werden. Spring Boot bietet vorkonfigurierte Einstellungen, wodurch sich der Aufwand für die Erstellung eines Projekts reduziert und Entwickler sich auf die Geschäftslogik konzentrieren können [3].

Der in Weiß abgebildete TCP-Server ist für die direkte

Kommunikation mit den E-Scootern zuständig. Nach einer initialen Konfiguration verbinden sich die IoT-Geräte der Fahrzeuge automatisch mit dem Server. Anschließend können Befehle wie „Sperren“, „Entsperren“ oder Standortabfragen über TCP gesendet und verarbeitet werden. Dazu wird ein Protokoll auf Anwendungsebene namens @Track Air Interface Protocol verwendet. Der Server verarbeitet eingehende Nachrichten, extrahiert relevante Informationen und leitet diese an die Datenbank zur Speicherung weiter. Die Sicherheitskonfiguration wird mittels Spring Security und Keycloak-Integration realisiert. Keycloak übernimmt die Benutzerverwaltung und Authentifizierung. Dadurch wird sichergestellt, dass nur autorisierte Benutzer Zugriff auf sensible Funktionen haben.
Über die REST-Controller werden Benutzeranfragen entgegengenommen. Diese ermöglichen z. B. das Starten oder Beenden einer Fahrt, das Abrufen von Scooter-Informationen oder das Senden von Befehlen zur sonstigen Konfiguration. Die API-Endpunkte sind durch Spring Security abgesichert und setzen eine Authentifizierung über Keycloak voraus.
Die Service-Klasse dient als zentrale Logikschicht. Hier werden REST- und TCP-Anfragen verarbeitet, Geschäftsregeln implementiert und Daten zwischen Controller, TCP-Server und Repository vermittelt.
Für die Persistierung der Daten kommt MongoDB zum Einsatz. Das Repository kommuniziert direkt mit der Datenbank und verwendet Spring Data MongoDB. Die Datenbank speichert Informationen zu Scootern, Buchungen und Geodaten.
Diese Geodaten ermöglichen die Nutzung des sogenannten Geofence-Mechanismus. Dieser bewirkt, dass Nutzer gewisse Regeln in bestimmten Zonen befolgen müssen. Es gibt eine Betriebszone, welche ein E-Scooter nicht verlassen darf, sonst wird er gesperrt und ist vorerst nicht nutzbar. Parkzonen geben vor, wo die Fahrzeuge nach einer Fahrt abgestellt werden dürfen und wo nicht.

Aufbereitung des Scooters

Normalerweise werden die Scooter zum Privatgebrauch in Verbindung mit einer App vom Hersteller angeboten [2]. Um das Gerät also mit einem neuen Server zu binden ist eine Änderung an den Netzwerkeinstellungen erforderlich. Ein Software Tool, das als *Connection Tool* bezeichnet wird, bietet die Möglichkeit diese und weitere Änderungen an der Konfiguration per Kabelverbindung zu tätigen. So können einerseits grundlegende Parameter wie der Startmodus oder die Maximalgeschwindigkeit geändert werden, aber auch Echtzeitoperationen, wie das Entsperren bzw. Sperren oder Ein- und Ausschalten des Scheinwerfers, vorgenommen werden.

Verfügt der Scooter über eine SIM-Karte mit gültigem Mobilfunknetz-Vertrag, so verbindet sich das netzwerkfähige Gerät nach einer gewissen Zeit mit dem vorgegebenen Server, solange dieser über eine öffentliche IP-Adresse bzw. URL erreichbar ist. An diesen werden dann sämtliche Informationen gesendet, wie zum Beispiel intervallweise Nachrichten, die zur Statusabfrage dienen.
Darüber hinaus ist der Server die einzige Kommunikation-Schnittstelle für den Scooter und somit auch derjenige, der Befehle zur Laufzeit sendet und eingehende Bestätigungen erwartet. Damit können die zuvor beschriebenen Echtzeitaktionen über Nachrichten im Funknetz durchgeführt werden. Über diese Einbindung der Funktionen des TCP-Servers in die Spring Boot Anwendung, können durch den Client ausgelöste Anfragen, Befehle an ein bestimmtes Fahrzeug gesendet werden, wie in Abbildung 2 in dem gelben Teil dargestellt. Umgekehrt kann der Client durch bestimmte Anfragen, Informationen des Scooters erhalten, welche dieser zuvor als Bericht an den Server gesendet hat und somit in der Datenbank abgelegt wurden (siehe grüner Teil in Abb. 2).

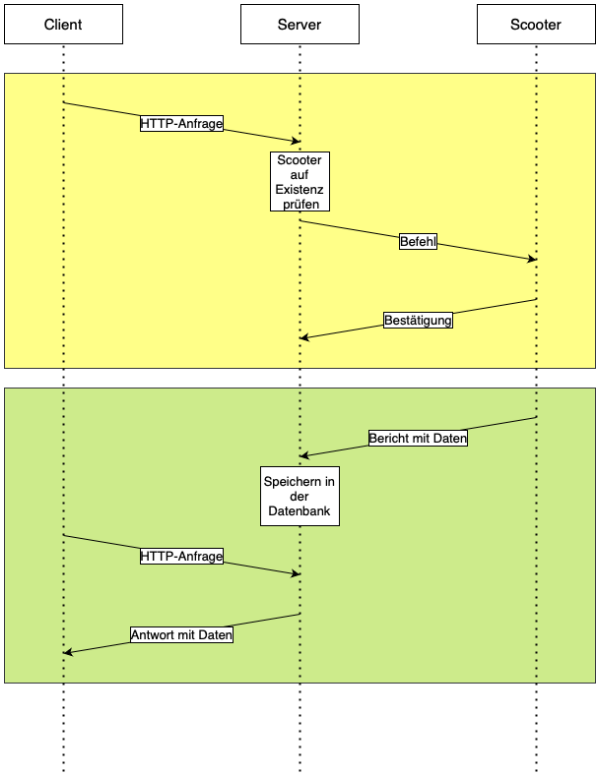


Abb. 2: Sequenzdiagramm [1]

Ausblick

Das Backend-System kann im weiteren Verlauf des Projekts noch durch neue Funktionen erweitert wer-

den. Die E-Scooter bieten eine große Auswahl an Konfigurationsmöglichkeiten um sie für den Nutzen zu optimieren. Zum Beispiel haben die Fahrzeuge eine Menge an Sensoren, die in der Umsetzung dieser Arbeit zwar nicht essenziell sind, im weiteren Verlauf aber interessant werden könnten. Alle weiteren Funktionen für Nutzer sowie Betreiber benötigen eine entsprechende Backend-Schnittstelle.
In der Anwendung sind zusätzlich für die Nutzer zugängliche Gameification-Funktionen geplant. Diese machen die Verwendung der App für Studenten attraktiver, da zum Beispiel durch gesammelte Punkte, Rabatte auf die nächsten Buchungen bevorstehen können. Auch diese Additionen benötigen entsprechende Backend-Funktionalitäten.
Für eine grafische Nutzung ist eine Frontend-App geplant, welche die im Backend implementierten Funktionen grafisch für den Nutzer in einer mobilen App darstellt. Benutzer sollen, wie in bekannten Scooter-Flotten-Apps auf einer Karte die Standorte verfügbarer Fahrzeuge einsehen können und dann mit dem angelegten Nutzerkonto eine Buchung durch Klicken des Objekts durchführen können.
Dieses Frontend dient also als Client zur Backend-App und schickt die erforderlichen Anfragen um entsprechende Funktionen auszulösen.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Christian Schwannecke. Okai Global - Was kann die App? https://e-roller.com/e-scooter/okai-global-was-kann-die-app/, 09 2022.

[3] Patrik T. Building REST API Using Spring Boot: A Comprehensive Guide. https://medium.com/@pratik.941/building-rest-api-using-spring-boot-a-comprehensive-guide-3e9b6d7a8951, 06 2024.

Kostenanalyse und Performance-Benchmarking der Stammdaten-Transformation nach S/4 Hana unter Nutzung der Stammdatenplattform Stibo STEP

Jannik Woeste **Thomas Rodach**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Robert Bosch GmbH, Feuerbach

Motivation

Die digitale Transformation stellt Unternehmen vor weitreichende Herausforderungen, insbesondere bei der Ablösung oder Migration bestehender IT-Systeme im Datenmanagement. Die Einführung von SAP S/4HANA zählt zu den bedeutendsten Vorhaben für Firmen, die SAP als zentrales ERP-System einsetzen. Ein kritischer Bestandteil dieser Transformation ist die Migration und Qualitätssicherung der vorhandenen Stammdaten. Fehlerhafte oder unvollständig migrierte Daten können zu erheblichen Mehrkosten führen und die Geschäftsabläufe empfindlich stören. [3] Um die Risiken zu minimieren, setzen viele Unternehmen ergänzend auf spezialisierte Stammdatenplattformen wie Stibo STEP. Sie bieten die Chance, die Daten vor der Übernahme zu bereinigen und zu harmonisieren, sodass die Migration nicht nur als technischer Umzug, sondern auch als Möglichkeit zur Optimierung der Datenbasis genutzt wird. Dies ist insbesondere vor dem Hintergrund der großen Ziele einer S/4HANA-Transformation relevant: Häufig stehen die Verbesserung von Geschäftsprozessen, die Modernisierung der Systemlandschaft und die Einführung innovativer digitaler Geschäftsmodelle im Vordergrund. [2]

Wie Abb. 1 verdeutlicht, sind die wichtigsten Treiber für die Transformation auf SAP S/4HANA die angestrebte Prozessverbesserung (65 % der Unternehmen) und die Modernisierung der ERP-Landschaft (55 %) – dicht gefolgt von der Unterstützung neuer Geschäftsmodelle und der Verbesserung der Datenverfügbarkeit in Echtzeit. Kostensenkung allein wird zwar genannt, ist aber nur für 16 % der Befragten der Hauptgrund. Dennoch müssen bei einer S/4-Umstellung die Kosten und Nutzen genau abgewogen werden, da eine Migration mit erheblichen Investitionen verbunden ist.

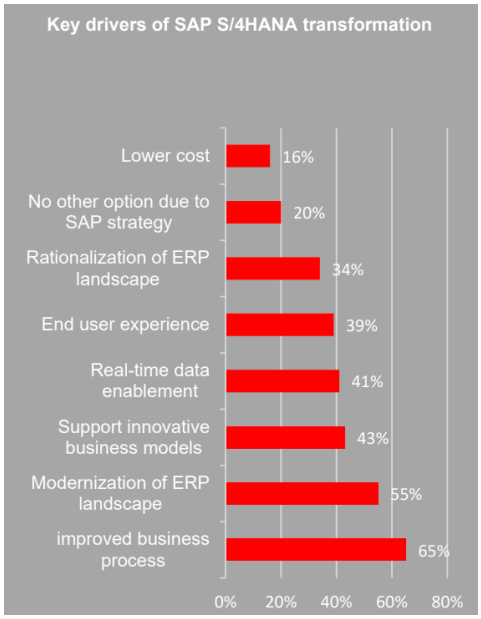


Abb. 1: Hauptgründe für die Transformation auf SAP S/4HANA. Studie von LeanIX und PwC [2]

Zielsetzung

Ziel dieser Bachelorarbeit ist es, nicht nur die Kosten der Stammdaten-Transformation zu analysieren, sondern durch den Einsatz der Master-Data-Management-Plattform Stibo STEP zu bewerten. Mit der Stammdaten-Transformation mit Stibo STEP sollen langfristige Prozessverbesserungen, geringere Betriebskosten und höhere Datenqualität realisiert werden. Diese Arbeit untersucht, inwiefern sich diese Ziele technisch und wirtschaftlich erreichbar sind.

Vorgehen

Aus Gründen der Vertraulichkeit wurden sämtliche Zahlenwerte in dieser Arbeit entfernt. Die Migration wurde in vier Phasen gegliedert: Analyse, Bereinigung, Übertrag und Validierung. Zunächst wurden Dubletten und Formatprobleme in Stibo STEP identifiziert und behoben. Das System diente dabei als Zwischenspeicher zur Bildung sogenannter "Golden Records".

Eine wirtschaftliche Bewertung erfolgte über eine Total Cost of Ownership-Analyse (TCO). In die Kalkulation flossen direkte Projektkosten (Kosten die bei der Beschaffung und beim Betrieb entstehen), indirekte Kosten (Kosten die durch Produktivitätsverlust oder Ausfallzeiten entstehen) [1] und Betriebskosteneinsparungen ein. Abb. 2 zeigt gut, was alles unter direkte, und was unter indirekte Kosten fällt.

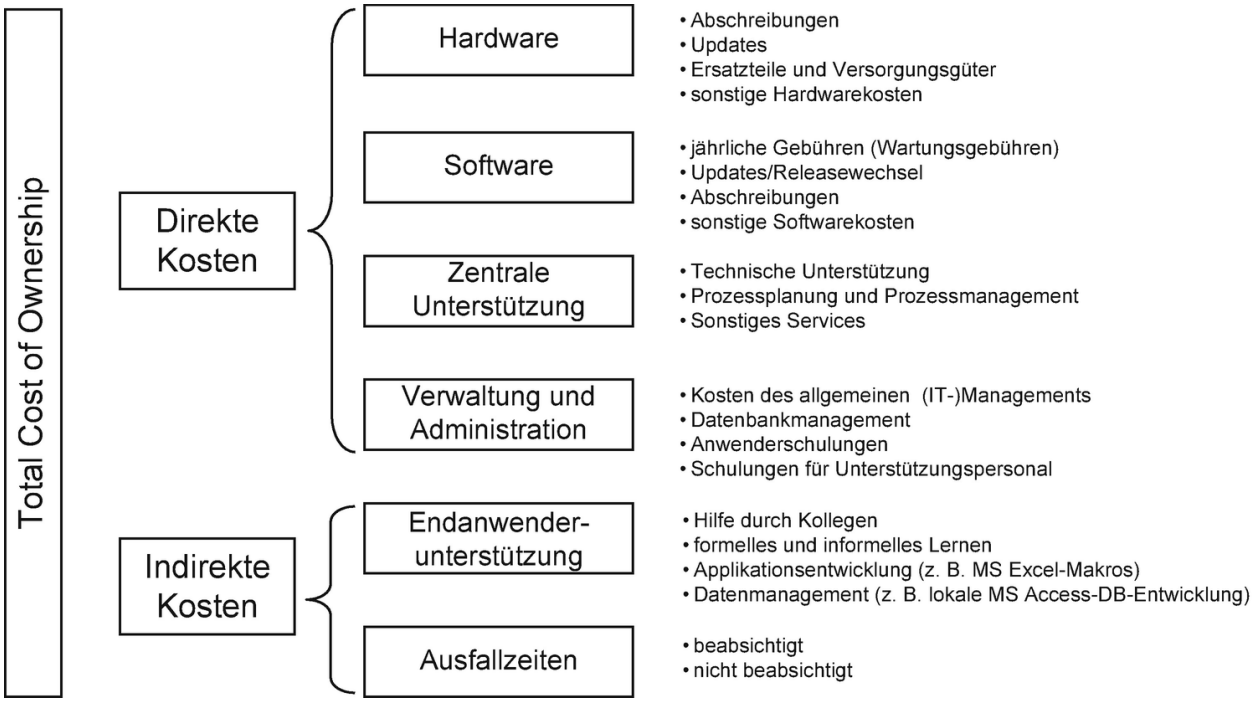


Abb. 2: Direkte und Indirekte Kosten in einer TCO-Analyse nach Müller, Lang und Hes [1]

Zusätzlich zur TCO wurde auch eine Break-Even-Analyse durchgeführt, um den Zeitpunkt zu bestimmen, an dem sich die Investition amortisiert. Parallel zur monetären Betrachtung wurde ein Performance-Benchmarking durchgeführt. Dabei handelt es sich um einen systematischen Leistungsvergleich verschiedener Migrationsansätze, in diesem Fall mit und ohne Einsatz von Stibo STEP. Ziel war es, Kennzahlen zur Effizienz der Migration zu erheben. Beispielsweise die Dauer der Datenübernahme, die Datenmengen pro Zeiteinheit, die Anzahl aufgetretener Fehler oder notwendiger Nacharbeiten usw. Diese Größen erlauben eine objektive Beurteilung der technischen Leistungsfähigkeit und fließen wiederum in die Bewertung der Wirtschaftlichkeit ein. So lässt sich nachvollziehen, inwiefern die zusätzliche MDM-Plattform den Prozess beschleunigt oder qualitativ verbessert und ob dies die damit verbundenen Kosten rechtfertigt.

Ausblick

Die Ergebnisse der Analyse zeigen, dass sich der Einsatz von Stibo STEP im Rahmen einer SAP S/4HANA-Migration insbesondere bei großen Datenmengen als sinnvoll erwiesen hat. Die durchgeführten TCO- und Break-Even-Analysen belegen die Wirtschaftlichkeit dieser Lösung unter den gegebenen Rahmenbedingungen.

Literatur und Abbildungen

[1] Andreas Gadatsch. *IT-Controlling*. Springer Fachmedien Wiesbaden, 2021.

[2] Stephan Kerner. The state of SAP S/4HANA Transformation. https://www.pwc.de/de/strategie-organisation-prozesse-systeme/the-state-of-sap-s4-hana-transformation.pdf, 2018.

[3] Luc Majewski. Herausforderungen bei Kernbankmigrationsprojekten. https://crossconsulting.de/herausforderungen-bei-kernbankmigrationsprojekten/, 06 2024.

Optimierung der Rechnungsverarbeitung durch Generative AI (GenAI): Automatisierte Analyse und Lösung von Buchungsfehlern

Bengue Yalcinkaya **Thomas Rodach**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Robert Bosch GmbH, Stuttgart

Einleitung

Die digitale Transformation stellt Unternehmen vor die Herausforderung, ihre Prozesse kontinuierlich zu optimieren und innovative Technologien gezielt einzusetzen. Insbesondere im Rechnungswesen gewinnen intelligente Automatisierungslösungen an Bedeutung, um komplexe Anforderungen effizient zu bewältigen. Fehler in der Rechnungsverarbeitung, vor allem im steuerlichen Kontext, führen häufig zu erhöhtem manuellem Aufwand und bergen Risiken hinsichtlich der Einhaltung regulatorischer Vorgaben. Diese Arbeit untersucht, wie Robotic Process Automation (RPA) und Generative Artificial Intelligence (GenAI) kombiniert eingesetzt werden können, um steuerliche Buchungsfehler automatisiert zu erkennen, zu analysieren und zu beheben. Während RPA die strukturierte Ausführung regelbasierter Prozesse übernimmt, unterstützt GenAI bei der Dateninterpretation und der Generierung von Lösungsvorschlägen. Anhand eines praxisorientierten Proof of Concept (PoC) bei der Robert Bosch GmbH wird das Potenzial dieser hybriden Automatisierungslösung demonstriert und hinsichtlich ihrer Wirkung auf Effizienz, Prozessqualität und operative Entlastung analysiert.

Problemstellung und Zielsetzung

Trotz fortschreitender Digitalisierung bleibt die Rechnungsverarbeitung fehleranfällig, insbesondere bei steuerlichen Buchungen. Unvollständige oder falsche Steuerkennzeichnungen führen häufig zu Buchungsfehler, deren manuelle Korrektur aufwendig und fehleranfällig ist. Ziel dieser Arbeit ist es, eine Automatisierungslösung zu entwickeln, die RPA und GenAI verbindet. Diese soll steuerliche Buchungsfehler automatisch erkennen, analysieren und beheben.

Theoretische Grundlagen

Prozessautomatisierung mit RPA

RPA übernimmt strukturierte und wiederkehrende Tätigkeiten, die sonst manuell ausgeführt würden, und bietet dabei eine höhere Ausführungsgeschwindigkeit sowie geringere Fehleranfälligkeit. [4]

Abb. 1: Roadmap zur RPA-Implementierung (eigene Darstellung in Anlehnung an Herm) [3]

Die Einführung von einer RPA beginnt mit der Bedarfsermittlung und der Auswahl einer geeigneten Softwarelösung, gefolgt von einem PoC zur Bewertung von Umsetzbarkeit und Nutzen, bevor der unternehmensweite Wissenstransfer den Gesamtprozess abrundet.

GenAI als Innovationstreiber in der Finanzabwicklung

Durch den gezielten Einsatz von GenAI lassen sich Prozesse in der Rechnungsverarbeitung deutlich verbessern. Ein wesentlicher Vorteil liegt in der Automatisierung der Datenerfassung aus Rechnungen, wodurch manuelle Eingaben und damit verbundene Fehlerquellen deutlich reduziert werden können. Charakteristisch für diese Modelle ist ihre ausgeprägte Anpassungsfähigkeit an unterschiedliche Eingabeformate. Durch gezieltes Training können sie so konfiguriert werden, dass sie mit einer Vielzahl von Rechnungsformaten umgehen. Darüber hinaus ermöglicht ihr Einsatz eine automatisierte

Kategorisierung von Dokumenteninhalten, wodurch eine strukturierte und systematische Weiterverarbeitung gewährleistet wird. Die damit einhergehende Automatisierung standardisierter Arbeitsschritte entlastet das Personal und schafft Kapazitäten für komplexere, wertschöpfende Tätigkeiten. [1]

Intelligent Automation im Finanzbereich: RPA und GenAI im Zusammenspiel

Der Fortschritt in der Prozessautomatisierung macht die begrenzte Anwendbarkeit klassischer RPA deutlich, insbesondere bei komplexen Prozessen mit unstrukturierten Abläufen oder hohem fachlichen Wissensbedarf. In solchen Fällen zeigt sich, dass RPA-Lösungen ohne zusätzliche kognitive Fähigkeiten nur eingeschränkt einsetzbar sind. Fortschritte im Bereich GenAI ermöglichen in diesem Kontext eine Weiterentwicklung hin zur Intelligent Automation (IA). Die Integration klassischer Prozessautomatisierung mit lernfähigen, adaptiven Systemen erweitert das Spektrum automatisierbarer Anwendungsbereiche. Dadurch können künftig auch unstrukturierte oder komplexe Prozesse unterstützt werden, die über rein regelbasierte Abläufe hinausgehen. Im Wesentlichen bedeutet das: Durch den Einsatz von IA wird die reine Prozessautomatisierung erweitert, sodass auch komplexere Aufgaben bewältigt werden können, die bisherige RPA-Lösungen überfordert hätten. [5]

Methodik

Die Arbeit verfolgt einen anwendungsorientierten Ansatz, bei dem theoretische Grundlagen mit einem praktischen Umsetzungsteil kombiniert werden. Im ersten Schritt erfolgt eine strukturierte Analyse der bestehenden Rechnungsverarbeitungsprozesses. Die Abbildung zwei zeigt den Ablauf der steuerlichen Verarbeitung von Rechnungen in einem internen Verwaltungssystem. Die Auswahl der Steuerkennzeichen richtet sich nach dem zugrunde liegenden Steuerregelwerk und dem jeweiligen Rechnungstyp. Jedem Steuerkennzeichen ist ein spezifischer Steuersatz zugeordnet. Diese Informationen werden zur Buchung an das SAP-System übermittelt. Bei fehlerhaften Steuerkennzeichnungen wird eine Buchungsfehlermeldung generiert.

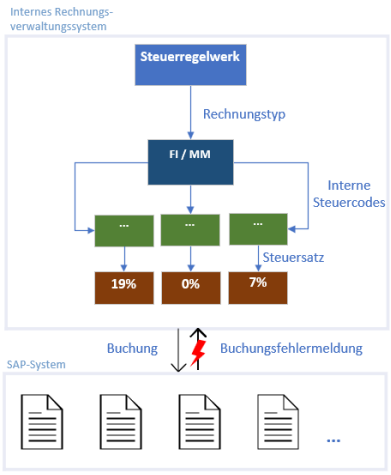


Abb. 2: Ablauf der Rechnungsverarbeitung im internen System und die Fehlermeldung vom SAP-System [2]

Darauf aufbauend wird ein Konzept zur Integration von RPA und GenAI entwickelt. Die prototypische Umsetzung erfolgt anhand realer Prozess -und Rechnungsdaten des Unternehmens. Abschließend wird die Lösung anhand definierter Kriterien hinsichtlich ihrer Funktionsfähigkeit und ihres Potenzials zur Effizienzsteigerung und Fehlerreduktion evaluiert.

Ausblick

Der entwickelte PoC dient als Ausgangspunkt zur Untersuchung des Potenzials fortschrittlicher Automatisierungstechnologien in der systemgestützten Fehlerbehandlung steuerlicher Buchungen. In der aktuellen Phase wird die Lösung exemplarisch in einem einzelnen Buchungskreis getestet, um erste Erkenntnisse über Umsetzbarkeit, Leistungsfähigkeit und Herausforderungen zu gewinnen. In Zukunft soll die entwickelte Lösung auch auf weitere Buchungskreise übertragen werden. Dadurch kann überprüft werden, inwiefern sie skalierbar ist und sich in unterschiedliche Systemlandschaften integrieren lässt. Außerdem soll bewertet werden, welchen praktischen Nutzen sie im betrieblichen Umfeld entfalten kann. Die Arbeit liefert somit eine Grundlage für die Weiterentwicklung intelligenter und datengetriebener Prozessautomatisierung im Finanzwesen.

Literatur und Abbildungen

[1] Abhijeet Singh Bais and Navneet Sharma. Generative AI Document Extraction Using Composite Approach: An External Data Integration. In *The AI Revolution: Driving Business Innovation and Research*, volume 525, page 405. Awwad, Bahaa, 2024.

[2] Eigene Darstellung.

[3] Lukas-Valentin Herm et al. A Consolidated Framework for Implementing Robotic Process Automation Projects. In *Business Process Management 18th International Conference, BPM 2020, Seville, Spain, September 13–18, 2020, Proceedings*, page 476. Fahland, Dirk; Ghidini, Chiara; Becker, Jörg; Dumas, Marlon, 2020.

[4] Xavier Lhuer. The next acronym you need to know about: RPA (robotic process automation). *McKinsey & Company*, page 1, 2016.

[5] Laura Minkova et al. From Words to Workflows: Automating Business Processes. https://arxiv.org/pdf/2412.03446, 12 2024.

HOCHSCHULE ESSLINGEN

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. Film & Media Christian Haas

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