AI5production - SYMPOSIUM

Digitalisierung in der Produktion - Lösungen und Anwendungen

Ergebnisse aus drei Jahren EDIH AI5production

15. OKTOBER 2025

JKU LINZ, OIC STUFENFORUM

Co-funded by the European Union

European Digital Innovation Hubs Network

Bundesministerium Wirtschaft, Energie und Tourismus

FFG Forschung wirkt.

Herausgeber

Johannes Kepler Universität Linz
Altenberger Straße 69
4040 Linz

Redaktionelle Leitung
Univ.-Prof.in Dr.in Barbara Krumay, Bakk MSc(WU)
barbara.krumay@jku.at

Für die inhaltliche Richtigkeit zeichnen die Autor:innen der Beiträge verantwortlich!

Linz, am 15.10.2025

Vorwort

Vorwort Dipl.-Ing. Franz Sümecz
TU Wien Pilotfabrik, Konsortialführer EDIH AI5production

Die schwache Konjunktur, die US-Handelspolitik und die Dekarbonisierung der Industrie stellen produzierende Unternehmen in Europa und im speziellen in Österreich vor enorme Herausforderungen. Es geht für die Unternehmen darum wettbewerbsfähig und vor allem resilient auf einem globalen Markt zu sein. Die Digitalisierung und vor allem KI spielen dabei eine zentrale Rolle. Österreichische Universitäten und außeruniversitäre Forschungseinrichtungen bewegen sich mit ihren Forschungs- und Innovationsleistungen im internationalen Spitzenfeld. Ziel des EDIHs ist, es den Technologietransfer zu forcieren, um das vorhandene Know-how und die vorhandene Expertise in die Anwendung bei den produzierenden Unternehmen zu bringen, um so die österreichische Wirtschaft nachhaltig zu unterstützen. Digitale Technologien ermöglichen eine Effizienzsteigerung in den Produktionsprozessen und Produktionsabläufen. Es geht darum Kosten zu senken, den Umsatz zu steigern, wettbewerbsfähiger und resilienter zu werden. Wie Digitalisierung und KI dazu beitragen, das zeigt der EDIH AI5production. Mit mehr als 150 Projekten mit Unternehmen in den letzten drei Jahren, wovon 50% KI-Methoden zum Inhalt hatten, ist der EDIH eine österreichische Erfolgsgeschichte. Wesentlich ist die einfache, niederschwellige und schnelle Unterstützung, die der EDIH den Unternehmen bietet und die große Expertise des Konsortiums bestehend aus 17 Institutionen. Für dieses Symposium wurden Projekte ausgewählt, die das Potential der Implementierung von digitalen Technologien und KI veranschaulichen und so auch anderen Unternehmen einen Anreiz bieten sollen die Möglichkeiten und Services des EDIHs zu nutzen, um die Digitalisierung und die Anwendung von KI in ihren Betrieben zu forcieren.

Univ.-Prof.in Dr.in Barbara Krumay, Bakk. MSc(WU)
JKU Business School, Leiterin Arbeitspaket „Support to find Investment"

Mit dem EDIH AI5production ist es gelungen, Digitalisierung in KMUs und Mid-Caps in Österreich voranzutreiben. Die vielfältigen Beiträge, die von den Partnern des EDIH AI5production eingereicht wurden, zeigen das eindrucksvoll. Dabei sind die Projekte, die die Grundlage für die Beiträge sind, nur der Anfang. Im Mittelpunkt stehen nicht Einzellösungen, sondern die Generalisierbarkeit der Erkenntnisse als Beitrag zu Wissenschaft und Wirtschaft. So entstehen praxisnahe Beispiele („Use Cases"), die auf wissenschaftlicher Basis erarbeitet wurden und als Anregung für die weitere Digitalisierungsvorhaben dienen. Insbesondere für die Nutzung und Implementierung von KI-Anwendungen wird der Ruf nach gut dokumentierten Use Cases immer lauter, um Unternehmen vor Fehlinvestitionen zu bewahren und die Potentiale der KI weitgehend nutzen zu können. Die in diesem Band zusammengefassten Beispiele sind Teil des AI5production-Symposiums, das am 15. Oktober 2025 an der JKU im OIC Stufenforum stattfindet. Drei der Beiträge beschäftigen sich mit der Rolle des EDIH AI5production für die Steigerung der Digitalen Reife von KMUs und Mid-Caps in der EU. Die restlichen Beiträge zeigen anschaulich und auf unterschiedlichen technischen Niveaus, wie Digitalisierung in produzierenden Unternehmen umgesetzt werden kann. Auch wenn KI in vielen Beiträgen vorkommt und in vielen TBIs eingesetzt wurde, dürfen auch andere, etablierte digitale Technologien nicht vergessen werden, um die europäischen und österreichischen KMUs und Mid-Caps in der Digitalen Reife voranzubringen.

Inhaltsverzeichnis

Digitale Reife von Kleinen und Mittleren Unternehmen durch Innovationsimpulse 3
Schulungen und Unternehmenskultur als Faktoren für KI-Integration 9
Nutzen für produzierende Unternehmen durch das Europäische Innovationsökosystem 14
Künstliche Intelligenz in der Produktion 21
Digitale Datendurchgängigkeit für die Antriebstechnik in der Gebäudetechnik 29
Effizienzsteigerungen bei WIG- und Plasma-Schweißinvertern 33
Multi-Sensor-Imaging für die qualitativen Charakterisierung von gebrauchten Solarmodulen 39
Digitalisierung eines Energieeffizienzkonzepts mit Reportingfunktion 42
Gießprozess für Sandformen mit dynamischer Vakuumsteuerung ROBOCAST V 47
CFD-Simulation als Digitales Service für Entwicklung und Produktion 51
Intelligente Chatbots zur Abfrage von Unternehmenswissen - ALLEX 58
Optimierung von Umformprozessen durch KI-basierte Überwachung und Steuerung 65
Digitale mensch-zentrierte Assistenzsysteme 73
Intelligentes Assistenzsystem für flexible Mensch-Roboter-Kollaboration in der Fertigung 79
Open Source Cybersecurity Anomalieerkennung 84
RAG is Not All You Need! 90
Rekonstruktionspipeline für Low-Cost Aufnahmen zur 96
Towards Optical Tracking for AR Assisted Oral Surgery 104

Glossar

- European Digital Innovation Hub (EDIH) AI5production – einer von vier in Österreich durch EU und FFG geförderten Hubs mit Fokus auf produzierende KMUs und Mid-Caps
- Nationale Digital Innovation Hubs (DIHs) – regionale Anlaufstelle zu Digitalisierungsfragen.
- Test before Invest (TBI) - Kleinprojekt zum Thema Digitalisierung, das ein Partner des EDIH AI5production für ein Unternehmen durchführt. Ein TBI inkludiert 2 Trainings (Technologietraining zu Beginn des Antrags und „Training am Objekt“ zu Projektende).
- Support to find Investment (STFI) - Services und Beratungen zur Finanzierung von Digitalisierungsprojekten im Rahmen des EDIH AI5production. Die Bandbreite reich von Beratung in Hinblick auf weitere Forschungsförderung sein (für den Fall, dass weitergehende Entwicklungen notwendig sind) bis hin zu Unterstützung bei der Suche Venture Capital oder Business Angels
- KMUs & Mid-Caps – entsprechend der Definition der Europäischen Union mit Fokus auf die Zahl der Mitarbeiter:innen
- Start-Ups – Unternehmen mit innovativen Geschäftsideen und -modellen bis max. 5 Jahre nach der Gründung
- Künstliche Intelligenz (KI) – allgemein als nicht-regelbasierte Technologie zu verstehen, die die Intelligenzleistung von Menschen nachahmt („mimic“) – weitere Definitionen und Kategorien sind in den Beiträgen zu finden

Digitale Reife von Kleinen und Mittleren Unternehmen durch Innovationsimpulse

Barbara Krumay
JKU Business School, Johannes Kepler Universität Linz
barbara.krumay@jku.at

Manuel Mühlburger
JKU Business School, Johannes Kepler Universität Linz
manuel.muehlburger@jku.at

Kurzzusammenfassung

Für kleine und mittlere Unternehmen (KMUs) ist die Umsetzung von Digitalisierungsprojekten und damit einhergehende die Hebung der Digitalen Reife herausfordernd. Es fehlen oft die finanziellen und personellen Möglichkeiten. Darüber hinaus müssen auch interne Voraussetzungen erfüllt werden, damit Digitalisierung möglich ist. In der Regel sind kleiner Unternehmen auch vorsichtig bei Investitionen in neue Technologie, daher fehlt oft eine Infrastruktur zum Testen neuer digitaler Technologien. Externe Innovationsimpulse sind für KMUs dann von großer Bedeutung, wenn sie leicht zugänglich und unabhängig (z.B. von Herstellern) sind. Basierend auf Interviews mit Vertreter:innen von Unternehmen, die die Services des EDIH AI5production in Anspruch genommen haben, wurde festgestellt, dass diese Innovationsimpulse zur Umsetzung von Digitalisierungsprojekten geführt haben und dadurch digitalen Reifegrad gesteigert werden konnte.

Keywords: Digitalisierung, Digitale Reife, Innovationsbereitschaft, Innovationsimpulse

Abstract

The implementation of digital technologies and digitalization projects is challenging, in particular for small and medium-sized enterprises (SMEs). They often lack the financial and human resources to implement such projects. In addition, digitalization projects also require the fulfillment of internal preconditions. Furthermore, SMEs are generally cautious about investing in new technology and often lack the infrastructure to test new digital technologies. External innovation impulses are of great importance for SMEs if they are easily accessible and independent (e.g., from manufacturers). Based on interviews with companies that have been consumed the services of the EDIH AI5production, it has been shown that these innovation impulses can contribute to implementing digitalization projects and raising the level of digital maturity.

Keywords: Digitalisation, digital maturity, Innovation readiness, innovation impulses

Einleitung

Kleine und mittlere Unternehmen stellen 99 % der europäischen Unternehmen dar und werden oft als Motor der europäischen Wirtschaft bezeichnet [1]. Die Europäische Union (EU) hat daher Initiativen gesetzt, um KMUs in ihren Digitalisierungsprojekten zu unterstützen und deren Digitale Reife zu heben [1]. Viele Herausforderungen lassen sich auch auf die Entwicklung von KMUs zurückführen, wie historisch gewachsenen Strukturen und unklar definierten Verantwortlichkeiten – vieles wird zur „Chefsache“ erhoben, unabhängig ob entsprechende Expertise vorhanden ist. Viele KMUs werden von den Gründer:innen oder Besitzer:innen geführt – häufig sind es auch Familienbetriebe [2], die sich dadurch auszeichnen, dass strategische Entscheidungen vor allem von den Gründer:innen und Familienmitgliedern getroffen werden. Diese Entscheidungsträger:innen haben Erfahrung im eigentlichen Kerngeschäft des Unternehmens, aber oft fehlt die Zeit, um technisches Know-How aufzubauen. Das wirkt sich auch auf die Digitalisierungsbereitschaft der Unternehmen aus. Familiengeführte Unternehmen zeigen auch oft geringe Risikobereitschaft [2], worunter vor allem die Innovationsfreude leidet. Governancestrukturen, die die Verantwortlichkeiten regeln, existieren im

eigentlichen Sinn nur implizit, d.h. das Wissen über die Verantwortung für Entscheidungen ist nicht ausdrücklich im Unternehmen definiert und dokumentiert [2].

KMUs werden in der EU über Mitarbeiterzahl (MA), den Umsatz und die Bilanzsumme definiert. Mittlere Unternehmen haben demgemäß max. 250 MA, einen Umsatz von max. 50 Mio € und eine Bilanzsumme von max. 43 Mio. €. Für kleine und Kleinstunternehmen sind Bilanzsumme und Umsatz gleich, d.h. kleine Unternehmen haben max. 50 MA und einen Umsatz von max. 2 Mio. €, Kleinstunternehmen max. 10 MA und einen Umsatz von 2 Mio. € [3]. Darüber hinaus definiert die EU sogenannte Mid-Caps mit bis zu 3.000 MA [4]. Durch die Größe von KMUs ergeben sich bestimmte Charakteristika (z.B. Ressourcenknappheit, fehlende Forschung und Entwicklung, durch operatives Geschäft getriebene Ziele, keine dokumentierte Unternehmensstrategie), die ihre Innovationskraft einschränken können. Unternehmen, in denen überwiegend Gründer:innen in entscheidungsrelevanten Positionen sitzen [5] sind in der Regel jüngere Unternehmen in der Start-up Phase. Als Start-up werden KMUs bezeichnet, deren Gründung maximal fünf Jahre zurückliegt und die durch innovative Technologie oder innovatives Geschäftsmodell ein Mitarbeiter- oder Umsatzwachstum aufweisen. In der Realität werden Start-ups meist von einer Einzelperson oder einem Gründer:innen-Team gegründet, die ein gemeinsames Ziel oder eine gemeinsame Idee haben. Neben Start-ups sind in der europäischen KMU-Landschaft viele Familienunternehmen vertreten, die - unabhängig von Größe und Bestehen des Unternehmens – bestimmte Charakteristika aufweisen. Dazu gehören, dass die Entscheidungen bei den Mitgliedern einer Familie liegen, die das Unternehmen gegründet hat, mindestens ein Familienmitglied in die Unternehmensführung eingebunden ist und mehr als ein Familienmitglied im Unternehmen eine entscheidungsrelevante Position innehat [7], [8], [9]. Ein-Personen-Unternehmen werden von der Betrachtung daher meist ausgeschlossen [7], [8]. Familienunternehmen zeichnen sich durch einzigartige personelle Ressourcen, bestimmte Strukturen und Nachfolgeregelungen aus [8]. Der Erfolg von Familienunternehmen kann durch die Beteiligung von Familienmitgliedern sowohl positiv – durch stärkere Identifikation mit dem Unternehmen – als auch negativ – durch die familiäre Situation und die finanzielle Abhängigkeit - beeinflusst werden [10]. Alle Unternehmen durchlaufen einen Reifungsprozess und damit einhergehend eine strukturelle Transformation. Diese kann sowohl durch äußere (z.B. Marksituation, neue Technologien) als auch durch innere Einflüsse (z.B. Wechsel an der Unternehmensspitze) angestoßen werden. Jede Transformation bringt Veränderungen mit sich, die von verschiedenen Faktoren (z.B. Unternehmenskultur, Änderungs- und Innovationsbereitschaft) abhängen. Die digitale Transformation von Unternehmen ist eine Transformation, die durch vorhanden universelle digitale Technologien und der Adoption im Markt getrieben wird. Um eine digitale Transformation von KMUs möglich zu machen, müssen digitale Fähigkeiten und Innovationsfähigkeit (als interne Faktoren) sowie eine adäquate Umgebung (externe Faktoren) entsprechend gegeben sein [6]. Damit KMUs diese Transformation schaffen, sind aufgrund von geringen personellen und monetären Ressourcen auch externe Impulse. Externe Innovationsimpulse können unterschiedlich gestaltet werden, wobei diese Angebote mögliches niederschwellig gestaltet werden müssen. Neben der Möglichkeit, den Einsatz neuer Technologien zu testen (TBI) zählen auch monetäre Instrumente (z.B. im Rahmen von STFI Förderungen, Kontakt zu Business Angels), Schulungsangebote und Vernetzungsmöglichkeiten zu diesen Impulsen. In diesem Beitrag wird untersucht, wie externe Innovationsimpulse zur Umsetzung von Digitalisierungsprojekten und somit zur Steigerung der digitalen Reife beitragen können.

Vorgehen

Um zu untersuchen, wie externe Innovationsimpulse die digitale Reife von KMUs steigern können, wurden mit drei Unternehmen, die im Rahmen des EDIH AI5production ein TBI durchgeführt haben, Interviews geführt. Die Unternehmen wurden bewusst so ausgewählt, dass sie sich anhand ihrer internen Faktoren unterscheiden, während der Innovationsimpuls durch das TBI immer annähernd gleich ist (Services des EDIH AI5production in Anspruch genommen, Steigerung der Digitalen Reife dokumentiert). Ausgewählt wurden ein Familienunternehmen (Unternehmen A), ein Start-Up (Unternehmen B) und ein Unternehmen, das bereits restrukturiert wurde (Unternehmen C), d.h. einen Transformationsprozess durchlaufen haben (siehe Tabelle 1). Im Zeitraum von Mai bis Juli 2025 wurden

leitfadengestützte Interviews mit mehreren Personen im Unternehmen geführt. Diese Interviews wurden aufgezeichnet, KI-gestützt transkribiert und qualitativ analysiert.

Tabelle 1. Zusammenfassende Darstellung der Unternehmen & Interviews

	Unternehmen A	**Unternehmen B**	**Unternehmen C**
Einordnung	Familienunternehmen	Start-Up	Restrukturiert
Alter in Jahren	~ 25	< 5	> 50
Anzahl MA	< 250	< 50	< 3.000
Geschäftsfeld	Metallverarbeitung	Zulieferer	Elektronik
Technologie	KI-Anwendung	Plattform	KI-Anwendung
Interviewpartner:innen	Geschäftsführer:in (A1) Produktionsleiter:in (A2)	Gründer:in (B1 & B2)	Geschäftsführer:in (C1) Innovationsmanager:in (C2) MA in der Produktion (C3)

Der Interviewleitfaden wurde aus der Literatur entwickelt und enthielt vier Leitfragen:

- Welche externe Innovationsimpulse brauchen KMUs, um Digitalisierung im Unternehmen umzusetzen? (Bedarf)
- Wie müssen externe Innovationsimpulse gestaltet sein, damit die Innovationsimpulse von den KMUs angenommen werden? (Gestaltung)
- Welche internen Faktoren beeinflussen den Erfolg von Innovationsimpulsen? (Interne Faktoren)
- Welche Beispiele im Unternehmen gibt es, wie externe Innovationsimpulse zur Steigerung der digitalen Reife beigetragen haben? (Beispiele)

Zusätzlich wurden kurze Einleitungsfragen (zum Unternehmen und zur Person) gestellt. Während des Interviews wurde auf Verständnisfragen der Interviewpartner:innen eingegangen. Es wurde bewusst darauf verzichtet, den EDIH AI5production im Leitfragen zu nennen. Die Interviews dauerten im Schnitt 47 Minuten.

Ergebnisse

Um abzubilden, wie externe Innovationsimpulse zur Steigerung der digitalen Reife von KMUs beitragen können, werden die Ergebnisse der Interviews anhand der Leitfragen dargestellt. Einige Aussagen von Interviewpartner:innen werden paraphrasiert dargestellt und mit dem Kürzel der Interviewpartner:in versehen.

Hinsichtlich der ersten Leitfrage „Welche externe Innovationsimpulse brauchen KMUs, um Digitalisierung im Unternehmen umzusetzen?“ wurden unterschiedlichste Möglichkeiten genannt, die in drei Kategorien zusammengefasst werden können: Monetäre Aspekte, Know-How-Aufbau und Zugang zu Technologien. Unter den monetären Innovationsimpulsen wurden vor allem Förderungen und der Zugang zu Investor:innen genannt. Die Interviewpartner:innen von Unternehmen B (Start-Up) wiesen auf die große, aber oft unübersichtliche Förderlandschaft in Österreich hin und nannten auch Business Angels (B2) als wichtigen Innovationsimpuls, da „Business Angels nur innovative Unternehmen unter ihre Fittiche nehmen“ (B2). Die Interviewpartner:innen A1 und A2 wiesen auf die starke Position der Hausbank hin, die ebenfalls durch entsprechende Kredite Innovationsimpulse setzen kann. Allerdings wurde auch gesagt, dass „es vor fünf Jahren noch leichter war, Kredite für innovative Projekte zu bekommen“ (A1). In der Kategorie Know-How wurden alle Innovationsimpulse subsummiert, die mit Trainings und Schulungen zu tun haben. Während im Unternehmen B vor allem „Selbstlernen durch Youtube“ vorherrscht, haben die Interviewpartner:innen der beiden anderen Unternehmen eine sehr klare Vorstellung über die Notwendigkeit von strukturierten Angeboten. Interviewpartner:in A2 meinte dazu, dass „kontinuierliche Schulungen zu neuen Technologien aber auch Auffrischung von Know-How zu bereits genutzten Technologien ist unbedingt notwendig“. Interviewparten C3 wies darauf hin, dass „nur durch Training-on-the-Job KI nicht bewältigbar sein

wird“. Als möglichen Formen wurden Hackathons, Bootcamps und Schulungen durch Chatbots genannt. Alle Interviewpartner:innen haben betont, dass die Schulung „Hand in Hand“ mit der Einführung von neuen Technologien gehen muss bzw. „idealerweise schon lange vor der Einführung damit begonnen werden kann“ (C1). Nur dann entsteht durch Schulungen ein Innovationsimpuls. Als weiterer externer Innovationsimpuls wurde auch der Zugang zu Technologie, die im Unternehmen nicht vorhanden ist, genannt. Insbesondere wurde darauf hingewiesen, dass für viele Technologien die hohen Investitionskosten nicht sinnvoll sind. Die Interviewpartner:innen der Unternehmen A und B haben besonders darauf hingewiesen, dass erst für Unternehmen ab einer gewissen Größer eigene Forschungs- und Entwicklungsbestrebungen über die normale Weiterentwicklung des Kerngeschäfts hinaus sinnvoll sind. Interviewpartner:in B1 sagte dazu: „Wir nutzen alle möglichen Zugänge zu Hackerspaces, Factories und ähnliches, um neue Technologien auszuprobieren – selbst können wir uns diese Infrastruktur nicht leisten“ (B1).

In allen Interviews wurde darauf hingewiesen, dass externe Innovationsimpulse eine wichtige Rolle für die Umsetzung von Digitalisierungsprojekten gespielt haben. Mehrfach wurde erwähnt, dass im Unternehmen manche Projekte nur durch externe Innovationsimpulse möglich waren. Es wurde auch direkt auf die Rolle des EDIH AI5production eingegangen und angemerkt, dass „wir dieses Projekt sicher nicht ohne den AI5production gestartet hätten“ (C3). TBIs wurden ausdrücklich als Initialzündung für ein späteres Projekt bezeichnet (A1, C2). Auch STFI wurde dezidiert als wichtige Voraussetzung und als entscheidendes Service für die Umsetzung genannt (B1, B2, C1). Das Service wurde als besonders wichtig und wertvoll eingestuft. Interessanterweise wurden auch die Digital Maturity Assessments angesprochen, die vor und nach jedem TBI durchgeführt werden müssen. Diese „geben einen guten Überblick, wo wir stehen“ (A2).

Hinsichtlich der Frage, wie externe Innovationsimpulse gestaltet sein müssen, um von den KMUs angenommen werden gab es eine eindeutige Antwort, die sich am besten mit dem Begriff „niederschwellig“ zusammenfassen lässt. Auffällig ist, dass die Antworten vor allem negativ konnotiert waren (z.B. „nicht kompliziert“, „keine langen Anträge“, „wenig Bürokratie“, „nicht versteckt und nur für Kenner“). Interviewpartner:in B2 meinte, dass „man als Start-up viele Vorteile hat und von vielen Serviceangeboten profitiert – andere KMUs wissen oft nichts davon“. Interviewpartner:in C2 meinte, „wenn man einmal in diesem Radl drin [sic] ist wird es leichter – hineinzukommen ist schwierig“ (C2). Die Interviewpartner:innen von Unternehmen an, bisher kaum oder wenig Erfahrung damit zu haben, aber ziemlich klare Vorstellungen, wie man externe Innovationsimpulse „auch für uns verfügbar machen“ kann (A1) - dazu gehören maßgeschneiderte Angebote („damit man sich nicht durch alle wühlen muss“) und kurzfristige Möglichkeiten („muss zeitlich in unser Geschäft passen“). Kosten wurden dahingehend angesprochen, dass externe Innovationsimpulse „gratis oder angemessen“ (B2) sein müssen, damit sie auch angenommen werden. Betont wurde auch, dass alle Maßnahmen „viel Aufwand bedeuten“, der sich auch in Ressourcenbindung zeigt und somit auch Kosten verursacht. Interviewpartner:in B2 meinte dazu, dass „man viel Zeit und Schweiß in einen Antrag steckt, der dann abgelehnt wird“. Damit einhergehend wurde von allen Interviewpartner:innen eine Art Einschätzung von Erfolgschancen genannt, wenn externe Innovationsimpulse (v.a. Förderungen, Zugang zu Technologie) von Unternehmen in Betracht gezogen werden. Ein weiterer wichtiger Faktor für die Akzeptanz von externen Innovationsimpulsen ist deren Unabhängigkeit. Von allen Interviewpartner:innen wurde geäußert, dass reine „Verkaufsveranstaltungen“ von Technologieanbietern nicht als Impuls sondern viel mehr als „aggressives Marketing“ (B2) verstanden und somit häufig auch abgelehnt werden.

Ebenfalls sehr klar wurde die Frage nach den internen Faktoren, die den Erfolg von externen Innovationsimpulsen ausmachen, beantwortet. Alle Interviewpartner:innen waren sich einig, dass eine offene Unternehmenskultur und die Unterstützung durch das Management zwei Grundvoraussetzungen sind. Weiters wurden die Digitalisierungsbereitschaft der Mitarbeiter:innen und deren aktuelles Wissen (A1, B1, B2, C1), eine flexible oder ambidextere (Zusammenspiel aus Stabilität und Agilität) Struktur (B1, C1, C2, C3) und das Abbauen von Ängsten hinsichtlich Veränderung (A1, B1, C2) genannt. Auf

unterschiedliche Art wurde auch die Entwicklung von Vorbildern („Champions“ / „Role Models“ / „Stewards“) angesprochen. Über diese Vorbilder kann ein gezielter und ehrlicher Erfahrungsaustausch angestoßen werden, der Klarheit über externe Innovationsimpulse und deren Nutzung bringt. Damit können auch Ängste und Unsicherheiten in Zeiten von Veränderungen beseitigt werden.

Als letzte Frage im Leitfaden wurde erhoben, welche Beispiele im Unternehmen existieren, wie externe Innovationsimpulse zur Steigerung der digitalen Reife beigetragen haben. Obwohl bewusst nicht nach den Services des EDIH AI5production gefragt wurde, nannten alle Interviewpartner:innen die Services (TBIs, STFI, Schulungen) als Beispiele. Hervorgehoben wurde die Niederschwelligkeit der Services, aber auch „das Verständnis aller am EDH AI5production beteiligten für die Herausforderungen im produzierenden Bereich“ (A1). Weiters wurden die individuelle und direkte Kommunikation sowie die umfassende Servicekultur genannt. Andere Beispiele wie die generelle Zusammenarbeit mit Forschungseinrichtungen (Universitäten, Fachhochschulen, Forschungsinstituten), die Leistung von unterschiedlichen Anbietern (Wirtschaftskammer, Austrian Wirtschaftsservice) und die Möglichkeiten „aus der Erfahrung anderer zu lernen“ (B2) wurden ebenfalls angesprochen.

Interessanterweise wurde in keinem Interview Vernetzung als Einflussfaktor genannt. Auf Nachfrage wurden vor allem Zeit sowie „Fit“ genannt, die einer intensiveren Vernetzung im Kontext von externen Innovationsimpulsen entgegenstehen. Der „Fit“ bezieht sich dabei einerseits auf Inhalte (Vernetzung zu einem bestimmten Thema) als auch auf das Unternehmen. Interviewpartner:in A2 meinte dazu „wir wissen schon, wer zu uns passt“ (A2). Allerdings wurde von zwei Interviewpartner:innen (B1, C2) auch der positive Effekt von inhaltlich gut gestalteten und entsprechend orchestrierten Netzwerkveranstaltungen genannt und als externer Innovationsimpuls identifiziert.

Zusätzlich zu den Leitfadenfragen wurde auch analysiert, ob die Einordnung der Unternehmen in Familienunternehmen, Start-up und restrukturiertes Unternehmen zu anderen Ergebnissen führen. Neben einer zählend-interpretativen Analyse wurden auch im Rahmen des Interviews vorkommende Emotionen zur Einschätzung genutzt. Die Nennungen wurden in 0 / und mit + - +++ zur Darstellung der Häufigkeit der Nennung unterteilt und in der Relation zueinander eingeschätzt und über die Interviewpartner:innen je Unternehmen gemittelt. Emotionalität zu einem bestimmen Thema wird mit * (kein Stern – keine Emotion, * - *** von geringe bis hohe Emotionalität) operationalisiert. Untersucht wurden die ersten drei Fragen des Leitfadens.

Tabelle 2 Zusammenfassende Darstellung der Unternehmen & Interviews

	Unternehmen A	**Unternehmen B**	**Unternehmen C**
Einordnung	**Familienunternehmen**	**Start-Up**	**Restrukturiert**
Bedarf			
Monetäre Aspekte	+++ / ***	+++ / *	+/ 0
Know-How Aufbau	++ / *	++ / **	+++ / *
Zugang zu Technologien	+++ / **	+++ / ***	+ / 0
Gestaltung			
Niederschwelligkeit	++ / ***	+++ / 0	++ / *
Interne Faktoren			
Offene Unternehmenskultur	+ / 0	++ / **	+++ / *
Unterstützung durch Management	++ / **	+++ / 0	+++ / ***

Zusammenfassung

Externe Innovationsimpulse werden überwiegend als positiv und notwendig gesehen, um die Digitalisierungsprojekte umzusetzen und dadurch die Digitale Reife von KMUs und Mid-Caps zu steigern. Neben monetären Aspekten wurden Know-How-Aufbau und Zugang zu Technologien als

wichtige Innovationsimpulse genannt. Diese müssen möglichst niederschwellig gestaltet sein, damit diese auch genutzt werden. Die Unternehmen selbst müssen Voraussetzung schaffen, damit die Innovationsimpulse sich auch in einer gesteigerten Digitalen Reife widerspiegeln. Positive Beispiele von externen Innovationsimpulsen zeigen, dass neben Niederschwelligkeit aber auch das Verständnis für die Situation von produzierenden Unternehmen eine wichtige Rolle spielt. Deutlich wurde auch, dass unabhängige Angebote herstellerspezifischen vorgezogen werden. Festgestellt wurde auch, dass die Interviewpartner:innen externe Innovationsimpulse, insbesondere die Services des EDIH AI5production, maßgeblich für die Umsetzung von Digitalisierungsprojekten sind. Vergleicht man die Ergebnisse hinsichtlich der Einordnung der Unternehmen zeigt sich, dass Unternehmen A und B viele Übereinstimmungen haben. Der Fokus liegt auf den monetären Aspekten und der Möglichkeit, auf Technologien zugegriffen werden kann. Entgegen der häufig geäußerten Besonderheit von Familienunternehmen, scheint sich das für Innovationsimpulse nicht zu bewahrheiten. Auffallend ist, dass bei monetären Aspekten die Emotionalität beim untersuchten Familienunternehmen sehr hoch war. Es wurde von beiden Interviewpartner sehr emotional über die nicht erfolgreichen Versuche eine Förderung zu erhalten berichtet. Unternehmen C - als Mid-Cap und größtes und ältestes Unternehmen in der Untersuchung – scheint in allen Aspekten sehr ausgeglichen zu sein. Emotionalität konnte bei Unternehmen C nur hinsichtlich Unterstützung des Managements identifiziert werden. Betrachtet man die Aussagen zu den Beispielen von erfolgreichen Innovationsimpulsen, könnte man ebenfalls ableiten, dass für Mid-Caps die Angebote anders designed sein sollten, um die Digitale Reife zu steigern. Allerdings können aus den drei untersuchten Unternehmen keine allgemeingültigen Aussagen abgeleitet werden.

Literatur

[1] Europäisches Parlament, „Kleine und mittlere Unternehmen | Kurzdarstellungen zur Europäischen Union | Europäisches Parlament“, Kurzdarstellung zur Europäischen Union. Zugegriffen: 5. August 2025. [Online]. Verfügbar unter: https://www.europarl.europa.eu/factsheets/de/sheet/63/kleine-und-mittlere-unternehmen

[2] M. R.W. Hiebl, „Risk aversion in the family business: the dark side of caution“, Journal of Business Strategy, Bd. 35, Nr. 5, S. 38–42, Sep. 2014, doi: 10.1108/JBS-09-2013-0087.

[3] „Kleine und mittlere Unternehmen - EUR-Lex“. Zugegriffen: 14. August 2025. [Online]. Verfügbar unter: https://eur-lex.europa.eu/DE/legal-content/glossary/small-and-medium-sized-enterprises.html

[4] „SME definition - European Commission“. Zugegriffen: 14. August 2025. [Online]. Verfügbar unter: https://single-market-economy.ec.europa.eu/smes/sme-fundamentals/sme-definition_en

[5] B. Hendricks, T. Howell, und C. Bingham, „How much do top management teams matter in founder-led firms?“, Strategic Management Journal, Bd. 40, Nr. 6, S. 959–986, Juni 2019, doi: 10.1002/smj.3006.

[6] C. F. A. Arranz, M. F. Arroyabe, N. Arranz, und J. C. F. de Arroyabe, „Digitalisation dynamics in SMEs: An approach from systems dynamics and artificial intelligence“, Technological Forecasting and Social Change, Bd. 196, S. 122880, Nov. 2023, doi: 10.1016/j.techfore.2023.122880.

[7] C. Döfrlinger, A. Dörflinger, K. Gavac, und B. Vogl, „Endbericht Familienunternehmen“, KMU Forschung Austria, Wien, 2013.

[8] J. Diéguez-Soto, P. López-Delgado, und A. Rojo-Ramírez, „Identifying and classifying family businesses“, Rev Manag Sci, Bd. 9, Nr. 3, S. 603–634, Juli 2015, doi: 10.1007/s11846-014-0128-6.

[9] Europäische Kommission, „Overview of family business relevant issues“, Europäische Kommission, 30-CE-0164021/00–51, 2009. [Online]. Verfügbar unter: http://ec.europa.eu/enterprise/ policies/sme/promoting-entrepreneurship/family-business/index_en.htm

[10] A. Stewart und M. A. Hitt, „Why Can’t a Family Business Be More Like a Nonfamily Business? Modes of Professionalization in Family Firms“, Family Business Review, Bd. 25, Nr. 1, 2012.

[11] B. Weber-Lewerenz, „Corporate digital responsibility (CDR) in construction engineering—ethical guidelines for the application of digital transformation and artificial intelligence (AI) in user practice“, SN Applied Sciences, Bd. 3, Nr. 10, 2021, doi: 10.1007/s42452-021-04776-1.

Schulungen und Unternehmenskultur als Faktoren für KI-Integration

Karin Koivisto
Pilotfabrik TU Wien – Industrie 4.0
karin.koivisto@tuwien.ac.at

Kurzzusammenfassung

Künstliche Intelligenz (KI) in der Produktion ist nicht nur eine technologische Herausforderung, sondern brauchen in vielen Fällen organisationale Veränderung um erfolgreich zu sein. Um KI in produzierenden Unternehmen erfolgreich einsetzen zu können, müssen die Mitarbeiter:innen entsprechend geschult werden. Neben der Vermittlung von technologischen Inhalten können diese Schulungen aber auch, eine organisationale Veränderung vorantreiben. Neben den technologischen Themen wie grundlegendes Verständnis von KI, praxisnahe Anwendungsbeispiele sowie der Aufbau von Datenkompetenz sollen Mitarbeitende auch befähigt werden, Prozesse ganzheitlich zu betrachten und Potenziale entlang der Wertschöpfungskette zu identifizieren. Die Ergebnisse aus dem EDIH AI5production haben gezeigt, dass praxisorientierte Anwendungsfälle und integrative Workshops hinaus Akzeptanz, Digitale Reife und Bereitschaft zum Wandel fördern. Die Unternehmenskultur muss sich dafür aber in Richtung offene Lern- und Fehlerkultur entwickeln und die aktive Partizipation der Mitarbeiter:innen fördern, um Ängste abzubauen. Durch KI-Schulungen können somit nicht nur technische Fähigkeiten geschult werden, sondern darüber hinaus aus die Unternehmenskultur verändert werden, um einen nachhaltigen KI-Einsatz möglich zu machen.

Keywords: Künstliche Intelligenz (KI), Produktion, Mitarbeiterschulung, Unternehmenskultur

Abstract

Artificial intelligence (AI) in production is not only a technological challenge, but in many cases also requires organisational change in order to be successful. AI implementation in manufacturing companies requires training for the employees. In addition to technological content, these training courses can also drive organisational change. Besides pure technological topics such as a basic understanding of AI, practical application examples and the development of data literacy, employees should also be empowered to take a holistic view of processes and identify potential along the value chain. The results from EDIH AI5production have shown that practice-oriented use cases and integrative workshops promote acceptance, digital maturity and willingness to change. However, the corporate culture must evolve towards an open learning and error culture and encourage active participation by employees in order to reduce fears. AI training can therefore not only teach technical skills, but also change the corporate culture to enable the sustainable use of AI.

Keywords: Artificial Intelligence (AI), production, employee training, corporate culture

Einleitung

Künstliche Intelligenz (KI) wird derzeit als nächste „General Purpose Technolog“, also allgemein einsetzbare Technologie mit breitem Anwendungsspektrum diskutiert. Produzierende Unternehmen sehen KI als den entscheidenden Unterschied für ihre Wettbewerbsfähigkeit. Für den sinnvollen Einsatz von KI in produzierenden Unternehmen entstehen neben technologischen auch organisationale Herausforderungen, denn diese Technologien müssen von den Mitarbeiter:innen verstanden und beherrscht werden können. Die größte Hürde ist für viele nicht technischer, sondern organisatorischer und kultureller Natur. Schulungen mit Fokus auf technologische Aspekte sind ein wichtiger Baustein, um die Mitarbeiter:innen frühzeitig zu befähigen, KI auch zu nutzen und dadurch die Akzeptanz für KI-Anwendungen zu steigern [1]. Weitere Faktoren, die die nutzenstiftende Einsatz von KI-Anwendungen

beeinflussen sind die Digitale Reife des Unternehmens und die individuelle Digitalisierungsbereitschaft der Mitarbeiter:innen [2], [3], die sich auch in der Unternehmenskultur abbilden [1], [4]. In diesem Beitrag soll gezeigt werden, wie Schulungen dazu beitragen können, die Voraussetzungen zu schaffen um KI-Anwendungen in produzierenden Unternehmen langfristig nutzen zu können.

Betrachtet man existierende KI-Schulungskonzepte, fokussieren diese vor allem technische Aspekte der KI-Anwendung. Den Schulungsteilnehmer:innen soll ein grundlegendes Verständnis für KI vermittelt werden. Vermittelt werden technisches Basiswissen und verschiedene Begrifflichkeiten und Definitionen (Was ist eigentlich KI, wie unterscheiden sich „Machine Learning“ und „Deep Learning“, was kann KI leisten – und was nicht?), um KI entsprechend einordnen zu können. Häufig werden Anwendungsbeispiele (z.B. aus der Produktion) genutzt, um durch die Verknüpfung von Grundkenntnissen mit einer Anwendung das Verständnis zu fördern. Technische Grundlagen inkludieren auch Datenkompetenz („Data Literacy“) [5], also die Fähigkeit Daten zu verstehen (durch Beobachtung, Analyse, Evaluierung, Reflexion) und zu nutzen (Integration, Navigation, Sammlung, Präsentation) [5]. In Hinblick auf KI-Anwendungen müssen Mitarbeiter:innen verstehen, wie die Daten für die KI-Nutzung (z.B. für Predictive Maintenance, Qualitätssicherung, Automatisierung) vorbereitet werden müssen und vor allem welche Ergebnisse diese Daten liefern können [5]. Datenkompetenz inkludiert aber auch die Fähigkeit verantwortlich mit den Daten umgehen zu können. In produzierenden Unternehmen müssen die Mitarbeiter:innen diese Datenkompetenz auf Prozesse anwenden können, in denen Daten genutzt werden. Um die Potenziale der KI identifizieren zu können, müssen auch die Prozesse verstanden werden, die zur Wertschöpfung des Unternehmens beitragen [6]. Diese Prozesse sind mit den Unternehmenszielen wie Effizienz und Effektivität aber auch Kundenorientierung abgestimmt, werden aber nur selten in den Schulungskonzepten direkt adressiert. Durch die Nutzung von praxisnahen Anwendungsfällen kann auch Prozessdenken zusätzlich zu den technischen Aspekten vermittelt werden. So können auch die aus dem KI-Einsatz hervorgehenden Veränderungen von Rollen und Abläufen an der Schnittstelle zwischen Menschen, Prozessen und Technologie geschult werden. Schulungen können auch dafür genutzt werden, Ängste abzubauen, zum Beispiel durch die Einbindung von „Role Models“ (z.B. IT-affine Mitarbeiter:innen, Data Analysts, Instandhalter:innen). Werden in den Schulungen Echtdaten (z.B. aus dem eigenen Unternehmen) verwendet, steigt auch die Sicherheit in Umgang mit den eigenen Daten und führt zu ersten Erfolgserlebnissen. Dadurch wird die Motivation gesteigert und die Akzeptanz gefördert.

Eine weitere Voraussetzung für einen erfolgreichen KI-Einsatz ist eine entsprechende Unternehmenskultur [1], [4]. In diesem Kontext spielen Rolle spielen Führungskräfte eine besondere Rolle. Sie trage aktiv dazu, dass Unsicherheiten beseitigt werden und eine Veränderung stattfinden kann. Unsicherheiten, die ausgeräumt werden müssen betreffen vor allem den möglichen Arbeitsplatzverlust durch KI aber auch den möglichen Verlust über die Kontrolle der eigenen Aufgabe. Die Unternehmenskultur sollte es auch erlauben, Mitarbeiter:innen aktiv in die KI-Strategie des Unternehmen einzubinden und KI-Projekte partizipativ mitgestalten zu können. Findet diese Einbindung früh statt, steigt auch die Akzeptanz [1]. Die Unternehmenskultur sollte auch integrative Ansätze, die unterschiedliche Personen mit unterschiedlichen Wissensständen einbeziehen, fördern, da diese zur Bewusstseinsbildung über die Prozesse und Zusammenhänge beitragen [1], [4]. Gerade in produzierenden Unternehmen, in denen Abteilungen oft in festen Strukturen arbeiten, wird durch KI eine abteilungs- und rollenübergreifende Zusammenarbeit notwendig, um Potenziale ausschöpfen zu können. Durch Schulung können die also nicht nur das technische Wissen von Mitarbeiter:innen gesteigert, sondern eine entsprechende Unternehmenskultur vorausgesetzt, auch Unsicherheiten beseitigt werden [4].

Vorgehen

Basis für diesen Beitrag bilden die aggregierten Daten von Unternehmen, die im Rahmen des EDIH AI5production (mindestens) ein TBI umgesetzt und entsprechende Schulungsmaßnahmen ergriffen haben. Als Basis dienen 151 TBIs (Stand Sommer 2025), die im Rahmen des Projekts einer standardisierten Messung der digitalen Reife (Digital Maturity Assessment – DMA) mittels eines Tools (Digital Maturity Assessment Tool) unterzogen wurden [7], [8]. Dadurch wird zu einem bestimmten

Zeitpunkt der aktuelle Stand der Digitalen Reife festgehalten um zu ermitteln, in welchen Bereichen das Unternehmen Unterstützung brauchen könnte. Die Messung erfolgt zu drei Zeitpunkten - vor dem TBI (DMA-T0), ein Jahr nach dem TBI (DMA-T1) und zu einem späteren Zeitpunkt (DMA-T2). Auch ohne Beteiligung eines EDIH kann das Tool OpenDMA von Unternehmen angewendet werden, um die eigene Digitale Reife zu klären [7]. Erfasst werden sechs Dimensionen, die in weitere Subdimensionen untergliedert. Gemessen wird in Fragenkatlogen, die sowohl binär (ja/nein) als auch Likert-skaliert (1 – 5) sein können. Erfasst werden folgende sechs Dimensionen:

(1) Digitalisierungsstrategie des Unternehmens - bezogen auf geplante oder getätigte Investitionen
(2) Bereitschaft für den digitalen Wandel - Einschätzung der aktuellen Nutzung digitaler Technologien inkl. Mainstream- und fortschrittlichere Technologien
(3) Auf den Menschen ausgerichtete Digitalisierung – bezogen darauf, wie die Mitarbeiter:innen für bzw. durch digitale Technologien qualifiziert, eingebunden und befähigt werden inkl. Verbesserung der Arbeitsbedingungen mit dem Ziel ihre Produktivität und Wohlbefinden zu steigern
(4) Datenverwaltung und -vernetztheit – Art und Organisation der digitalen Datenspeicherung innerhalb des Unternehmens und deren Nutzung für Geschäftszwecke unter Berücksichtigung von Datenschutz und Cybersicherheit
(5) Automatisierung und künstliche Intelligenz – untersucht das Ausmaß wie digitale Mittel Automatisierung und Erkenntnisgewinn in Geschäftsprozessen ermöglichen
(6) Grüne Digitalisierung – erfasst die Fähigkeit die Digitalisierung mit einem langfristigen Konzept durchzuführen unter Berücksichtigung von Nachhaltigkeitsaspekten

Der Erreichungsgrad je Dimension wird prozentual abgebildet und über die Dimensionen hinweg ein DMAScore gebildet. Bei mehr als der Hälfte der erfassten TBIs wurde bereits zwei Messungen (DMA-T1) durchgeführt. Dieser Beitrag betrachtet vor allem 46 TBIs, in denen Möglichkeiten des KI-Einsatzes getestet wurden. Darüber hinaus wurden im Rahmen des EDIH AI5production Schulungsmaßnahmen der Unternehmen dokumentiert und analysiert.

Ergebnisse

Der durchschnittliche digitale Reifegrad (DMAScore) lag bei der ersten Messung bei 51,4 % und stieg nach Durchführung eines TBI (DMA-T1) auf 59,5%. Bezieht man nur Unternehmen ein, bei denen zwei Messungen vorliegen (DMA-T0 und DMA-T1), stieg der durchschnittliche DMA um 9,3 % und in den einzelnen Dimensionen um 7.9 % (Digitalisierungsstrategie des Unternehmens), 8,0 % (Bereitschaft für den digitalen Wandel), 12,0 % (Auf den Menschen ausgerichtete Digitalisierung), 9,3 % (Datenverwaltung und -vernetztheit), 10,1 % (Automatisierung und künstliche Intelligenz) und 8,9 % (Grüne Digitalisierung). Einen Überblick über die Veränderungen in der DMA von Unternehmen mit zwei Messungen (DMA-T0 und DMA-T1) findet sich in Abbildung 1. Fokussiert man auf die TBIs, die die Einsatzmöglichkeiten von KI zum Inhalt hatten, liegen diese in allen Dimensionen über dem Schnitt. In der Dimension „Bereitschaft für den digitalen Wandel“ stieg das Ergebnis um 9 % und die Dimension „Datenverwaltung und -vernetztheit“ um 11%. Die höchste Steigerung zeigt sich mit 15 % in der Dimension „Auf den Menschen ausgerichtete Digitalisierung“. In dieser Dimension wurden spezifische Schulungspläne zur Umschulung und/oder Weiterbildung der Mitarbeiter:innen erfasst, die von den Unternehmen angeboten wurden. Zusätzlich wurden geplante Maßnahmen und Angebote - von regelmäßigen, umfassenden Schulungen zu fortschrittlichen Technologien oder zum Erwerb digitaler Kompetenzen – identifiziert. Diese sind auf die spezifischen Bedürfnisse und Schulungsanforderungen der Mitarbeiter:innen zugeschnitten. In den Unternehmen, die diese Maßnahmen umgesetzt haben, zeigt sich auch eine Steigerung der anderen Dimensionen. Im Detail sind interaktive Schulungen, mit weitreichenden Möglichkeiten Fragen zu stellen und unter Miteinbeziehung von Praxisbeispielen besonders erfolgversprechend. Im EDIH AI5production wurde auch deutlich, dass Workshops mit interdisziplinären Teams aus Fertigung, IT und Management die Akzeptanz von neuer Technologie (insbesondere KI) steigern, da Ängste abgebaut und konkrete Anwendungsfälle gemeinsam erarbeitet wurden. Durch die gemeinsame Erarbeitung entsteht auch ein stärkeres Verständnis für- und untereinander. Dafür ist eine offene Unternehmenskultur notwendig ist, die sich zum Beispiel in einer flachen Hierarchie, der Unterstützung durch das Management und gelebten integrativen Ansätze (z.B. in Form von Querschnittsgruppen) zeigt. Zusätzlich sind auch eine gelebte Fehlerkultur, die

Identifikation von „Role Models“ und klar definierte Unternehmengrundsätze zur KI-Nutzung wichtige Faktoren.

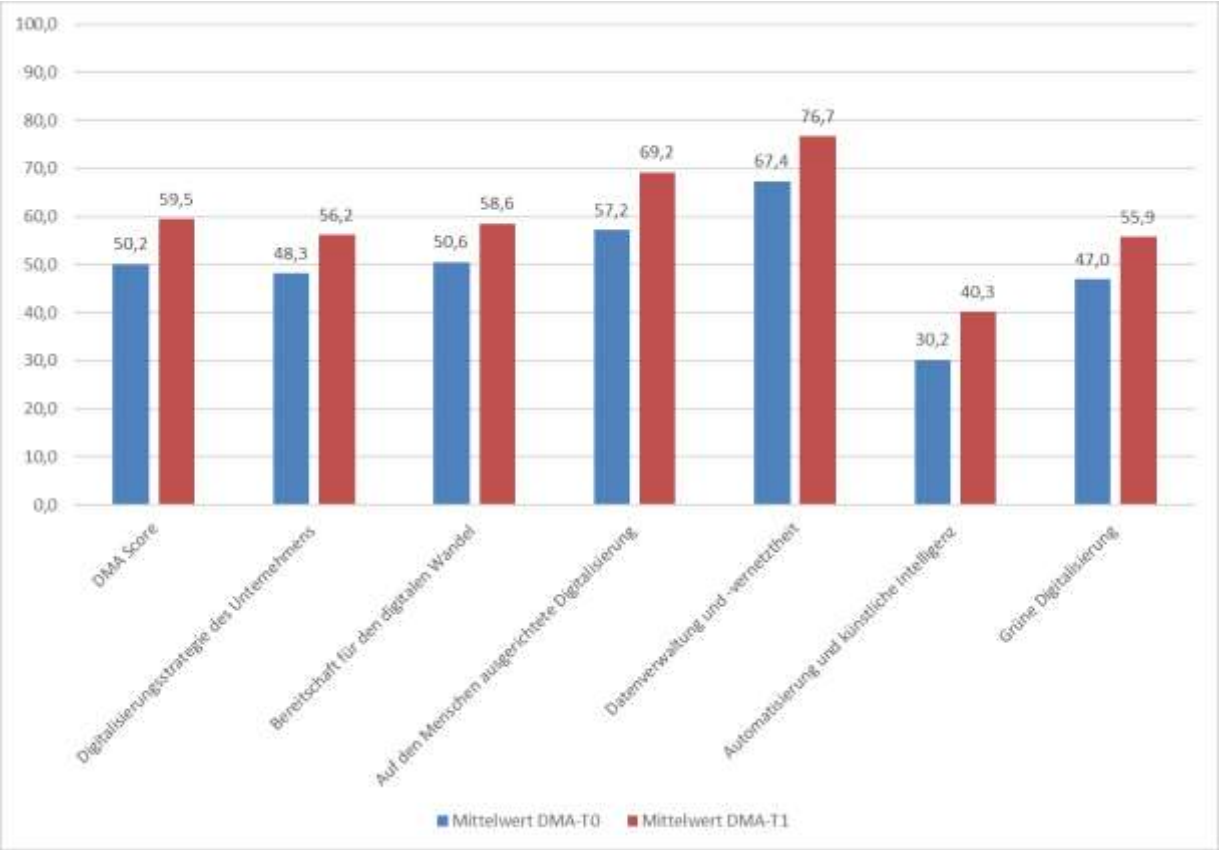

Abbildung 1. Veränderungen DMAScore und Dimensionen von Unternehmen mit mind. 2 Messwerten

Fazit und Handlungsempfehlungen

Um KI in produzierenden Unternehmen einzuführen und Unternehmensziele wie Steigerung von Effizienz und Effektivität aber auch Kundenzufriedenheit zu erreichen ist es notwendig, Mitarbeiter:innen frühzeitig diese neue Technologie nahezubringen. Schulungen decken dabei in der Regel vor allem die technischen Aspekte ab, aber durch Berücksichtigung organisationaler Aspekte, kann eine ganzheitliche Veränderung ermöglicht werden. Neben den technischen Grundlagen der KI müssen Aspekte wie Datenkompetenz und -sicherheit im Rahmen der Schulungen adressiert werden. Die Teilnehmer:innen sollen verstehen, dass sie Verantwortung für die Daten tragen, die sie in Unternehmensprozessen nutzen. Die Unternehmenskultur spielt vor allem dann eine Rolle, wenn in traditionellen Unternehmen in denen ein abteilungsübergreifendes Arbeiten nicht zwingend erforderlich war. Die Unternehmenskultur trägt auch zu einem wertschätzenden Umgang bei, der für die Zusammenarbeit im Kontext der KI notwendig ist. Um ein gemeinsames Verständnis zu schaffen, müssen Schulungen entsprechend gestaltet und koordiniert werden. Die Ergebnisse aus dem EDIH AI5production haben gezeigt, dass praxisnahe, interaktive und integrative Workshops hier eine wichtige Rolle spielen können. Um Ängste abzubauen und die Akzeptanz in die Technologie zu steigern, können Führungskräfte als Multiplikatoren fungieren. Sie sind auch dafür verantwortlich als Teil der Unternehmenskultur ein Klima zu schaffen, in dem Fragen und Zweifel ernst genommen werden.

Für Unternehmen, die KI einsetzen möchten, lassen sich aus den Ergebnissen des EDIH AI5production Handlungsempfehlungen ableiten. Für den KI-Einstieg empfiehlt es sich, zunächst zum Beispiel im Rahmen eines TBI mit integrierter Schulung die aktuelle Situation im Unternehmen zu analysieren. Dafür ist es sinnvoll, gemischte Teams aus allen Bereichen (Produktion, IT, Technik, Management) zusammenzubringen. So wird KI nicht als isolierte Technologie verstanden, sondern als integraler Bestandteil der Wertschöpfungskette. Zusätzlich empfiehlt es sich, „KI-Botschafter:innen“ zu

entwickeln, die als Ansprechpersonen dienen – ein Ansatz, der in TBIs zu einer wahrnehmbaren Verbesserung der Lern- und Innovationskultur führte. Die Einführung von KI steht immer in direktem Zusammenhang mit der Unternehmenskultur, die von Partizipation und Offenheit geprägt sein muss. Am Ende des Schulungsprogramms kann ein Team entstehen, das den Umgang mit Einsatzmöglichkeiten der KI nicht nur versteht, sondern auch kritisch hinterfragt und aktiv mitgestaltet. Die starke Praxisorientierung und der Fokus auf konkrete Anwendungsfälle machen dieses Konzept besonders geeignet für KMUs, die KI sinnvoll und ressourcenschonend einführen möchten. Ziel muss es sein, dass die Unternehmenskultur zu einer erfolgreichen KI-Implementierung beiträgt, indem eine positive Einstellung gegenüber Veränderung erzeugt wird und Akzeptanz gesteigert wird.

Literatur

[1] N. Rane, S. P. Choudhary, und J. Rane, „Acceptance of artificial intelligence: key factors, challenges, and implementation strategies“, Journal of Applied Artificial Intelligence, Bd. 5, Nr. 2, S. 50–70, Sep. 2024, doi: 10.48185/jaai.v5i2.1017.

[2] C. Williams, B. Krumay, D. Schallmo, und E. Scornavacca, „Digital maturity model for SMEs: validation through a mixed-method approach“, Pacific Asia Journal of the Association for Information Systems, Bd. 16, Nr. 1, S. 2, 2024.

[3] M. Muehlburger, B. Krumay, S. Koch, und S. Currle, „INDIVIDUAL DIGITAL TRANSFORMATION READINESS: CONCEPTUALISATION AND SCALE DEVELOPMENT“, Int. J. Innov. Mgt., Bd. 26, Nr. 03, S. 2240013, Apr. 2022, doi: 10.1142/S1363919622400138.

[4] O. T. Murire, „Artificial Intelligence and Its Role in Shaping Organizational Work Practices and Culture“, Administrative Sciences, Bd. 14, Nr. 12, S. 316, Dez. 2024, doi: 10.3390/admsci14120316.

[5] L. Van Audenhove, L. Vermeire, W. Van den Broeck, und A. Demeulenaere, „Data literacy in the new EU DigComp 2.2 framework how DigComp defines competences on artificial intelligence, internet of things and data“, Information and Learning Sciences, Bd. 125, Nr. 5–6, S. 406–436, Feb. 2024, doi: 10.1108/ILS-06-2023-0072.

[6] M. Mühlburger und B. Krumay, „Towards a context-sensitive conceptualisation of digital transformation“, Journal of Information Technology, Bd. 39, Nov. 2023, doi: 10.1177/02683962231219514.

[7] „Open DMAT - Digital Maturity Assessment tool | European Digital Innovation Hubs Network“. Zugegriffen: 5. Oktober 2025. [Online]. Verfügbar unter: https://european-digital-innovation-hubs.ec.europa.eu/open-dma

[8] F. Sümecz, C. Wögerer, und G. Hesina, „Der Weg zur digitalen Transformation in der Produktion, EDIH (European Digital Innovation Hubs Network)“, Technische Universität Wien, Pilotfabrik Industrie 4.0, 2025.

Nutzen für produzierende Unternehmen durch das Europäische Innovationsökosystem

Stefanie Werderits
Plattform Industrie 4.0 Österreich
stefanie.werderits@plattformindustrie.at

Roland Sommer
Plattform Industrie 4.0 Österreich
roland.sommer@plattformindustrie40.at

Kurzzusammenfassung

Der European Digital Innovation Hub (EDIH) AI5production bietet ein starkes, kooperatives Umfeld für KMUs und Mid-Caps – regional, national und europäisch – zur Förderung von Digitalisierung und Innovation. Strategische Auswahl von Partnerschaften mit anderen EDIHs und Stakeholdern, organisationsübergreifende sowie grenzüberschreitende Wissensvermittlung und gemeinsame Aktivitäten wie Webinare oder themenbezogene Netzwerk-Events ermöglichen den Austausch von Erfahrungen und Praxisbeispiele (Best Practices). Dadurch wird auch die Servicequalität der EDIHs verbessert. Die Erfahrungen, Herausforderungen und Potenziale im Aufbau von Netzwerken im Europäischen Innovationsökosystem aus dem Arbeitspaket „Innovationsökosystem und Netzwerken“ des EDIH AI5production werden in diesem Beitrag dargestellt, mit besonderem Fokus auf Mehrwert, Synergien und konkrete Unterstützungsangebote für Unternehmen.

Keywords: Netzwerk, EDIH, Innovationsökosystem, Kooperation, Digitalisierung

Abstract

The European Digital Innovation Hub (EDIH) AI5production offers a strong, collaborative environment for SMEs and mid-caps – regionally, nationally and across Europe – to promote digitalisation and innovation. Strategic partnerships with other EDIHs and stakeholders, cross-organisational and cross-border knowledge transfer, and joint activities such as webinars or themed networking events enable the exchange of experiences and best practices. This also improves the service quality of the EDIHs. This article presents the experiences, challenges and potential in building networks in the European innovation ecosystem from the ‘Innovation Ecosystem and Networks’ work package of the EDIH AI5production, with a particular focus on added value, synergies and concrete support services for companies.

Keywords: Networks, EDIH, innovation ecosystem, cooperation, digitalization

Einleitung

Digitale Transformation und Innovation gelingen nicht im Alleingang – sie erfordern funktionierende Netzwerke [1], [2]. Für viele Unternehmen ist der Zugang zu maßgeschneiderter technischer und strategischer Unterstützung im Digitalisierungsprozess aber schwierig. Es fehlt oft an Informationen über technologische Potenziale, verfügbare Förderangebote, passende Partner oder Beispiele aus der Praxis. Ein Ziel des EDIH AI5production ist es, ein starkes nationales und europäisches Innovationsökosystem für KMUs und Mid-Caps zu erschließen und weiterzuentwickeln. Eine Einbettung in ein funktionierendes Innovations-Ökosystem sowie existierende und neue Netzwerke kann dazu beitragen, die Innovationsentwicklungen von KMUs zu beschleunigen und Ressourcen bzw. Wissen zu kompensieren [3]. Innovationsökosysteme können als Systeme mit heterogenen, aber komplementären Akteuren definiert werden [4]. Diese sind nicht hierarchisch organisiert, sondern schaffen Werte durch modulare Strukturen und lose gekoppelte Kooperationen. Der EDIH AI5production tritt als orchestrierender Akteur im Sinne eines „Ecosystem Orchestrators“ auf – also einer Instanz, die Rahmenbedingungen und Schnittstellen etabliert [4]. Diese Rolle unterstreicht die Bedeutung von Komplementarität, Koordinationsmechanismen und gezielter Partnerintegration in

digitalen Innovationsnetzwerken [4]. Innovation in Ökosystemen entsteht durch lose gekoppelte, aber abgestimmte Akteure [4], die durch klar definierte Schnittstellen miteinander verbunden sind und komplementäre Leistungen (z.B. Beratung, technische Services und Finanzierung) anbieten. Netzwerke schaffen also Wert durch Zugang zu Wissen, Schnittstellen und Kooperation [4]. Zu den Herausforderungen der Zusammenarbeit in einem Ökosystem zählen Themen wie Abgrenzung, Überschneidung von Rollen, Zugang zu Ressourcen und Vertrauen [4]. Ziel ist es, die Elemente eines erfolgreichen Ökosystems im Rahmen von AI5production für den Nutzen von KMUs und Mid-Caps umzusetzen.

Vorgehen

Basierend auf den bereits vorgestellten Grundlagen von Ökosystemen wurde die Rolle des EDIH AI5production (insbesondere des Arbeitspaketes „Innovationsökosystem und Netzwerken“) identifiziert und für die Weiterentwicklung des Ökosystems verwendet. Zunächst wird die Ausgangssituation des AI5production dargestellt und danach eine konzeptionelle Einordnung der Rolle von AI5production vorgenommen.

Ausganssituation

Das EDIH-Netzwerk bestehend aus KMUs, Forschung, Politik und Intermediäre weist sowohl Modularität als auch Komplementarität. Das Netzwerk besteht es aus vielfältigen, oft dezentral agierenden Akteuren, die sich durch komplementäre Fähigkeiten ergänzen. Das Konsortium des EDIH AI5production besteht ebenfalls aus Partnern, die sich durch eine breitgefächerte Expertise ergänzen und erfüllt somit die Anforderungen an ein „Innovationsökosystems“. Darüber hinaus wird für die individualisierte Gestaltung der Zusammenarbeit und bei der gezielten Auswahl der Partner für EDIH-Kooperationen ebenfalls auf Modularität und Komplementarität geachtet. Konkrete Herausforderung im AI5production Projekt war bislang der Zugang zu Ressourcen durch die unterschiedliche nationale Förderausgestaltungen der EDIHs (z.B. nationale Vorgaben zur Ressourcenverwendung beschränkt auf KMUs im jeweiligen Nationalstaat) sowie Sprachbarrieren bei KMUs.

Dokumentation des Vorgehens

Kern der Strategie und zentrale Aufgabe des EDIH AI5production ist die Förderung und der Ausbau europäischer Netzwerke mit dem Ziel, KMUs und Mid-Caps durch den Zugang zum österreichischen und europäischen Innovationsökosystem und entsprechenden Networking-Möglichkeiten maßgeschneiderte Unterstützung zu bieten (siehe Abbildung 1). Die Vernetzungsstrategie umfasst folgende Aktivitäten:

- gezielte Auswahl von Partnern nach Themen- und Industriebezug
- Aufbau persönlicher Kontakte und formelle Zusammenarbeit via Memorandum of Unterstanding (MoUs)
- Entwicklung gemeinsamer Aktivitäten (z. B. Webinare, Dissemination-Events)
- Aktive Vermittlung relevanter Informationen und Services über das Netzwerk

Der Auf- und Ausbau des Ökosystems und der Netzwerk-Services ist hybrid gestaltet und kombiniert die strategische Planung, die hinter dem Netzwerkaufbau steht mit dynamischer Anpassung an Technologietrends und Bedarfe. Diese Bedarfe wurden auch immer wieder von den Unternehmen geäußert und umgesetzt. Ein anschauliches Beispiel dafür ist die Nachfrage nach Informationen zu einem möglichen Einsatz von Large Language Models (LLMs) in der Produktion. Dieser Bedarf wurde im Zuge der Aktivitäten von einer Reihe an Firmen geäußert und aufgenommen. Die Aktivitäten werden dokumentiert, evaluiert und laufend an neue Bedarfe angepasst.bei KMUs.

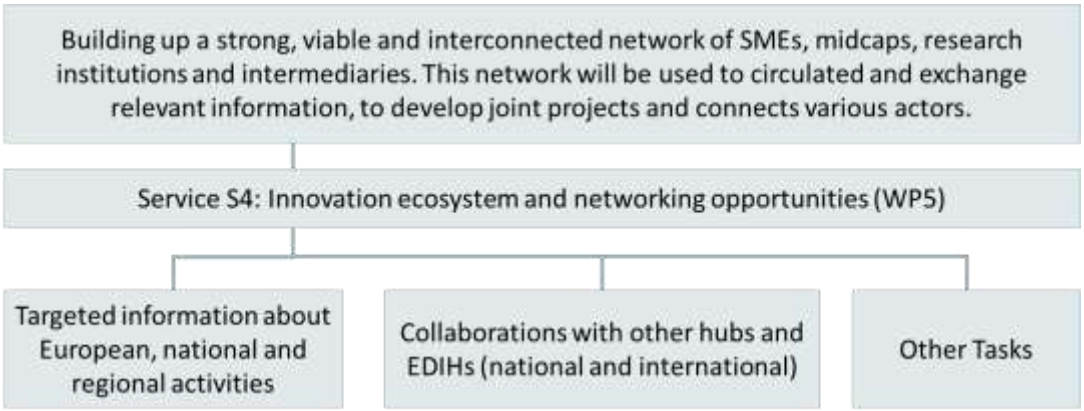

Abbildung 1. Ziele im Arbeitspaket Innovationsökosystem und Netzwerken

Die strategische Auswahl von Partnerschaften beruht auf bestimmten Faktoren wie Produktionsbezug aber auch thematische und geografische Nähe. Gleichzeitig tragen auch Partnerschaften, die sich im Laufe des Projekts organisch gebildet haben, zur Stärkung des Netzwerks bei. Die Aktivitäten im Rahmen der Kooperationen werden individuell gestaltet und sind auf die jeweiligen Fachkenntnisse und Rahmenbedingungen der einzelnen EDIHs zugeschnitten. Partnerschaften mit Stakeholdern außerhalb des Konsortiums werden über MoUs definiert, aber nicht hierarchisch gelenkt. Der Zugang zu Wissen, Schnittstellen und Kooperation erfolgt über Aktivitäten wie z. B. Webinare, Projektinitiativen oder die Vermittlung geeigneter Kooperationspartner. Diese „Wertschöpfung durch Verbindung“ ist ein wesentliches Ziel des EDIH AI5production. Die Integration in das europäische Ökosystem dient der Verbreitung relevanter Informationen, dem Austausch bewährter Verfahren, der Entwicklung gemeinsamer Projekte und der Vernetzung verschiedener Akteure (u. a. mit österreichischen KMUs und Mid-Caps).

Ergebnis

Es gibt eine Reihe an Ergebnissen aus dem Projekt AI5production, in dem ein funktionierendes Kooperationsnetzwerk in Österreich und Europa aufgebaut wurde.

Regionaler und nationaler Netzwerkaufbau

Auf regionaler Ebene sind die themenbezogenen Trainingsformate in Präsenz unter Mitwirkung unterschiedlicher Netzwerkpartner ein wichtiger Anknüpfungspunkt für Unternehmen, um in einem nächsten Schritt die Services des EDIH AI5production in Anspruch zu nehmen. Diese Formate, die über reine Disseminationsveranstaltungen (z.B. Stand bei Messe, Vortrag über das Projekt) hinausgehen, haben einen innovativen Charakter, da sie die Potenziale von AI5production und den Partnern gezielt nutzen und erfolgreiche Veranstaltungselemente verbinden. Zum einem gelingt es über das breite Partner-Netzwerk und dessen Kontakte, Forscher:innen mit Industrievertreter:innen zusammenzuführen und gleichzeitig Regionen Österreichs zu adressieren, welche oftmals von einer geringen Veranstaltungsdichte geprägt sind. Über gezielte Aktivierung der Netzwerke der Partner und die Einbindung relevanter Multiplikatoren, wie beispielsweise Interessensvertretungen, gelingt es die Aufmerksamkeit von KMUs und Mid-Caps zu gewinnen. Zum anderen hat die Kombination thematisch-fokussierter Information mit der Möglichkeit zum direkten Austausch mit Expert:innen und Gleichgesinnten positive Resonanz gezeigt. Um lokale KMUs anzusprechen, werden diese Veranstaltungen bewusst Vorort in unterschiedlichen Regionen in Österreich durchgeführt.

Die Grundidee hinter diesem Veranstaltungsformat ist, dass aktuelle Technologien verständlich aufbereitet und vermittelt werden. Forscher:innen übernehmen diese Aufgabe im Format, , während die Industrie anhand konkreter Beispiele („Use Cases“) reale Industrieanwendungen zeigt. Die Vertreter:innen der Hubs (EDIH, DIH) treten als „Honest Broker“ auf, das heißt – und das wird zu Beginn der Veranstaltungen auch klar vermittelt – es geht weder um eine Produktpräsentation noch um

jegliche Art von Kommerzialisierungsabsicht, sondern vielmehr um die Vermittlung von gesicherten, „neutralen“ Informationen. Flankiert werden die Veranstaltungen von einer Aufbereitung weiterführender österreichischer Förderprogramme, einem Aufzeigen von Pilotfabriken und Forschungszentren zu diesen spezifischen Themen und der Auflistung von Ansprechpersonen. Ein zentrales Element ist das Ermöglichen von intensiven Diskussionen zwischen den Teilnehmenden und den Vortragenden im Anschluss an den Vortragsteil. Nicht zuletzt tragen Erfahrungen aufgrund durchgeführter TBIs dazu bei, aktuelle Themen für die Veranstaltungen zu identifizieren. Die Qualität der Veranstaltungen ist somit auf das einzigartige Netzwerk der Partner und dessen gezielten Ausbau sowie auf das direkte Feed-Back im Austausch mit den Teilnehmern sowie den nahtlosen Übergang zu anderen Services zurückzuführen. Themen, die adressiert wurden, waren der industrielle Datenaustausch, der Digitale Produktpass, generative Künstliche Intelligenz und betriebliches Energiemanagement.

Ein weiterer Schwerpunkt war die gezielte Einbeziehung unterschiedlicher industrieller Netzwerke in Österreich, beispielsweise im Maschinenbau und der Elektro- und Elektronikindustrie. Mit diesen Akteuren konnte ein kontinuierlicher Wissenstransfer zwischen dem EDIH AI5production und seinen Netzwerkpartnern etabliert werden. Zur besseren Orientierung im Innovationsökosystem präsentiert AI5production auf seiner Website relevante Leuchtturmprojekte, Regularien und weiterführende Initiativen. Das Ziel dahinter ist, Ergebnisse aus national oder europäisch finanzierten Leuchtturmprojekten KMU-gerecht aufzubereiten. Vielfach werden in Leuchtturmprojekten konkrete Use Cases bearbeitet, deren Ergebnisse für andere Unternehmen von Relevanz sein können. Die in diesen Projekten adressierten Themen werden auch bei Veranstaltungen und Schulungen vertieft behandelt. Dank der Unterstützung der Netzwerkpartner können beispielsweise in regelmäßigen Abständen regionale Schulungen angeboten werden, die sich in ausgewählten Beiträgen aus Theorie und Praxis mit aktuellen Vorschriften und Technologien befassen.

Die wechselseitige Einbindung der Netzwerkpartner bei der Aussendung und Bewerbung von Veranstaltungen, sowie bei der Präsentation von Use-Cases, stärkt bei KMUs und Mid-Cap-Unternehmen in ganz Österreich das Bewusstsein für die Angebote der EDIHs. Der Ansatz „informieren, partizipieren und ausprobieren”, der durch diese Komponente des Kooperationsnetzwerks vorangetrieben wird, ist daher ein Erfolgsfaktor, um KMUs zu erreichen und sie mit Partnern zu verbinden.

Das europäische Netzwerk – zunehmende Zusammenarbeit der EDIHs

Der EDIH AI5production hat eine Reihe an europäischen Partnerschaften aufgebaut, um durch Erfahrungsaustausch, Veranstaltungen und gezielte Weiterbildungen den Zugang zu digitalen Technologien und Anwendungswissen zu verbessern und Synergien zwischen den EDIHs zu heben. Der europäische Netzwerkaufbau wurde dabei durch „Memoranda of Understanding‘“ unterstützt. Es gab einen intensiven Austausch zwischen den EDIHs zu unterschiedlichen Dienstleistungen, welche sie anbieten. Dies war insbesondere in der Anfangsphase des EDIHs eine Hauptaktivität, um voneinander zu lernen und gegebenenfalls Aktivitäten zu synchronisieren. Dabei standen neben Austausch zu Dienstleistungen auch andere Fragen im Fokus, wie beispielsweise bezüglich der zu adressierenden Zielgruppen, Abrechnungsmodalitäten, Reporting-Anforderungen oder Erfahrungen mit Zwischenevaluierungen.

Ein weiterer Schwerpunkt war der Zugang zu industriellen Best Practices und themenspezifischer Expertise für KMUs und Mid-Caps, auch über Ländergrenzen hinweg. Gemeinsam mit einem oder mehreren EDIHs organisierte Webinare haben sich als besonders effektiv und praktikabel in der Umsetzung erwiesen. Der Vorteil dieser transnationalen Webinare besteht darin, dass sie Unternehmen einen niederschwelligen Zugang zu Expertenwissen und inspirierenden Praxisbeispielen ermöglichen, sodass sie von den spezifischen Stärken und Erfahrungen unterschiedlicher Hubs profitieren können. Diese Formate bieten KMUs und Mid-Caps europäisches Know-how über Technologien, bewährte Unternehmenspraktiken und erfolgreiche Anwendungsfälle (z.B. erfolgreich durchgeführte TBIs),

welche aktuelle Themen, darunter Digitale Zwillinge, Security und Künstliche Intelligenz aufgegriffen hatten.

Die Gestaltung erster Webinare mit europäischen Partner EDIHs basierte auf der gemeinsamen Identifikation relevanter Themen und ist im Zuge individueller Gespräche mit den EDIHs entstanden. Auch die Idee, mehrere Hubs gemeinsam in ein thematisches Webinar miteinzubeziehen, um so Einblicke in unterschiedliche Best-Practices bzw. Zugang zu Experten zu erhalten, ist einem Ideenaustausch der EDIHs untereinander entsprungen. Somit zeigt sich das Potenzial kollaborativer Innovation der EDIHs. Ein physisches Pendant und Erfolgsbeispiel ist der Tschechisch-Österreichischer EDIH Twister in Brno [5].

Darüber hinaus hat das Veranstaltungsformat eines „EDIH-Workshop“, bei dem EDIHs sich Vorort in Wien getroffenen haben, die Möglichkeit zur Ideengenerierung bezüglich gemeinsamer Services geboten. Um auf ähnliche Weisen niederschwelligen Austausch zwischen den EDIHs zu ermöglichen und von ihren Einblicken in die Bedürfnisse der KMUs und Mid-Caps zu profitieren, wurde die Idee themen-spezifischer online Workshops mit den Partner-EDIHs ins Leben gerufen. So ist ein Webinar und Workshop zum Thema Digital Product Passport (DPP) in Planung. Dieser wird gemeinsam in Kooperation mit dem EU-Projekt Cirpass-2 organisiert und umgesetzt [6]. In Pilotierung in kleiner Runde soll getestet werden, ob dieses Format zur Ausrollung des DPP beitragen kann und inwieweit die Rolle der EDIHs beitragen kann, KMUs und Mid-Caps bei der Ausrollung des DPP zu unterstützen. Eine gezielte Reflektion der Rolle und Möglichkeiten der EDIHs bei der Unterstützung zur Implementierung von Technologien und Regularien kann folglich Unternehmen in ganz Europa zusätzliche Hilfestellung bieten.

Kooperationen der EDIHs fördern auch den Austausch von Informationen über aktuelle Entwicklungen, wie beispielsweise die Bedürfnisse von Unternehmen, Informationen über weitere Projekte, Vorschriften und Initiativen, die für KMUs und Mid-Cap-Unternehmen in der EU relevant sein können. Das EDIH-Netzwerk eignet sich für eine effektive Verbreitung von Informationen und bietet beispielsweise die Möglichkeit, das kürzlich gestartete „SECURE-Projekt“ [7] zur Umsetzung des „Cyber Resilience Act“ an die Zielgruppe zu kommunizieren.

Oftmals ist das Budget der EDIH-Kooperationspartner ausschließlich oder überwiegend für nationale Unternehmen reserviert (basierend aus Erfahrungen aus dem EDIH AI5production), wodurch die Möglichkeiten der gegenseitigen Unterstützung eingeschränkt werden – d. h. nicht alle EDIHs dürfen ihre Dienste EU-weit für KMUs und Mid-Caps zugänglich machen. Dennoch liefert der Austausch von Best Practices, beispielsweise zu den Aktivitäten und der Ausgestaltung der Dienstleistungen der EDIHs oder zu erfolgreichen Praxisbeispielen, wertvolle Erkenntnisse, die zur Qualitätsverbesserung des EDIHs AI5production beitragen.

Die Zusammenarbeit mit anderen EDIHs kann für den Zugang zu bestimmten Dienstleistungen, insbesondere Schulungen und TBIs, aber auch im Bereich der Finanzierung, einen Mehrwert bieten – sofern dies gemäß den nationalen Fördervereinbarungen möglich ist. Die Kooperationen der EDIHs ermöglichen den KMUs und Mid-Caps, über einen lokalen Hub in Kontakt mit einem breiten Netzwerk zu treten und Zugang zu Fachwissen der darin vertretenen Partner zu erhalten sowie potenziell von Dienstleistungsangeboten über die österreichischen Grenzen hinaus zu profitieren. Auf diese Weise kann innerhalb des Netzwerks die am besten geeignete Unterstützung durch einen Partner mit spezifischer Expertise ermittelt werden. Sprachbarrieren und geografische Entfernungen sollten jedoch als Hindernisse nicht außer Acht gelassen werden. In der Praxis konnten KMUs und Mid-Caps daher bislang nur begrenzt von bestimmten Dienstleistungen eines anderen europäischen Hubs Gebrauch machen.

Überregionales Netzwerk

Neben dem Aufbau eines europaweiten Ökosystems liegt ein weiterer Schwerpunkt auf der Schaffung eines nahtlosen Netzwerks in den österreichischen Grenzregionen und auf der überregionalen Zusammenarbeit. Es besteht eine Zusammenarbeit mit EDIHs in der Slowakei, Slowenien, der Tschechischen Republik, Deutschland und Ungarn. Auch mit zwei EDIHs in der Ukraine ist die Zusammenarbeit im Aufbau. Innerhalb Österreichs ist der EDIH AI5production gut mit den anderen österreichischen EDIHs und den nationalen DIHs (Digital Innovation Hubs) vernetzt.

Zusammenfassung

Unternehmen stehen vor großen strukturellen Herausforderungen, die sich aus den immer schnelleren Innovationszyklen, den disruptiven Elementen neuer Technologien im Produktionsprozess, insbesondere der Künstlichen Intelligenz, der Anpassung der Organisationsstrukturen an die raschen Technologiezyklen und Automatisierungsanforderungen, der Notwendigkeit der Weiter- und Höherqualifizierung der Mitarbeiter:innen sowie der Notwendigkeit der Implementierung datenbasierter Geschäftsmodelle ergeben. Industrie 5.0 beschäftigt sich im Kern mit Menschzentriertheit und damit mit Organisationsstrukturen und Qualifizierung. Der Fokus auf Aspekte von Industrie 5.0 sind bewusst im Fokus des EDIH AI5production - ausgedrückt durch die Zahl „5“ im Titel.

Die im EDIH angebotenen Dienstleistungen bieten Unternehmen die Möglichkeit, umfassend, individuell und gleichzeitig niederschwellig strukturelle Maßnahmen zur digitalen Transformation zu setzen. Damit hebt sich das Format von anderen Unterstützungsmaßnahmen ab. Die Innovation der Lösung besteht darin, dass unterschiedliche Dienstleistungen individuell auf die Bedarfe und Bedürfnisse der Unternehmen zugeschnitten werden. Der EDIH AI5production hat die internen Prozesse so aufgesetzt, dass beispielsweise bei TBI-Anträgen Unternehmen innerhalb von max. 72 Stunden eine Rückmeldung erhalten und somit sehr zeitnahe ihre Aktivitäten starten können.

Erwarteter Nutzen für Unternehmen

Die teilnehmenden Unternehmen profitieren von einer besseren Orientierung im europäischen Innovationsumfeld, Zugang zu Informationen, Trainings und europaweiten Angeboten, einem fokussierten Ansprechpartnernetzwerk und sich daraus ergebenden Projektchancen. Die Aktivitäten im Arbeitspaket „Innovationsökosystem und Netzwerken“ tragen wesentlich dazu bei, dass KMUs und Mid-Caps den Nutzen der Services des EDIH (TBI, Trainingsangebote und STFI) für sich erkennen und leisten somit einen Beitrag zur digitalen Transformation von Unternehmen. Auch tragen diese Aktivitäten wesentlich dazu bei, dass der EDIH AI5production potenziell interessierte KMUs und Mid-Caps initial erreichen kann, indem gezielt bestehende Netzwerke in die Aktivitäten eingebunden werden. Durch Formate, die den Austausch zwischen Unternehmen unterschiedlicher Größen (z.B. Großunternehmen und KMUs) entlang der Lieferkette (u.a. Zulieferer) sowie Experten verschiedener Institutionen (Wissenschaft, Beratung etc.) fördern, kann Vernetzungen innerhalb einer Branche entstehen. Über den EDIH AI5production hinaus können die gewonnen Kontakte den Unternehmen zugutekommen. In Zukunft wird ein weiterer Ausbau der regionalen Aktivitäten und internationalen Partnerschaften, die Vertiefung des Austauschs und die Erprobung neuer Formen der Zusammenarbeit angestrebt. Auch ein stärkerer Fokus auf das Thema künstliche Intelligenz adressiert eine Entwicklung, die für alle Unternehmen Bedeutung hat. Ziel ist es, mit jeder Aktivität zu einem höheren Innovations- und Digitalisierungsgrad beizutragen und somit Mehrwert für KMUs und Mid-Cap-Unternehmen in Europa zu erzielen - wodurch Spill-Over Effekte entlang den Wertschöpfungsketten zu erwarten sind.

Das Netzwerk des EDIH AI5production zeigt: wenn europäische Kooperation nicht als Pflicht, sondern als kreative Chance verstanden wird, entsteht ein belastbares Innovationsökosystem. Die nächsten Schritte die Zusammenarbeit zu heben, bestehen im Ausbau thematischer Schwerpunktaktivitäten (z.B. KI, Kreislaufwirtschaft durch Digitalisierung, Cyber-Security), der Identifikation von Praxisbeispielen und EDIH-übergreifenden Angeboten, der Bereitstellung eines Überblicks bestehender Angebote sowie

gezielter Adressierung aktueller Herausforderungen. Darüber hinaus ist die Vertiefung regionaler Netzwerke und die gezielte Ansprache der KMUs und Mid-Caps vor Ort ein wesentlicher Erfolgsfaktor für ein belastbares Innovationsökosystem.

Literaturverzeichnis

[1] A. Bargoni, A. Ferraris, Š. Vilamová, und W. M. H. Wan Hussain, „Digitalisation and Internationalisation in SMEs: a Systematic Review and Research Agenda“, Journal of Enterprise Information Management, Bd. 37, Nr. 5, S. 1418–1457, 2024.

[2] „European Innovation Ecosystems - Research and innovation“. Zugegriffen: 7. August 2025. [Online]. Verfügbar unter: https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-europe/european-innovation-ecosystems_en

[3] F. P. Hilmersson und M. Hilmersson, „Networking to accelerate the pace of SME innovations“, Journal of Innovation & Knowledge, Bd. 6, Nr. 1, S. 43–49, 2021.

[4] M. G. Jacobides, C. Cennamo, und A. Gawer, „Towards a theory of ecosystems“, Strategic Management Journal, Bd. 39, Nr. 8, S. 2255–2276, Aug. 2018, doi: 10.1002/smj.2904.

[5] „AI5production Erfolgsstories - Networking“, EDIH AI5production. Zugegriffen: 7. August 2025. [Online]. Verfügbar unter: https://ai5production.at/success-story/network/

[6] Platform Industrie 4.0, „Exploring the Role of EDIHs in Supporting SMEs with the DPP“. Zugegriffen: 30. September 2025. [Online]. Verfügbar unter: https://plattformindustrie40.at/blog/event/exploring-the-role-of-edihs-in-supporting-smes-with-the-dpp/

[7] Secure, „SECURE Strengthening EU SMEs Cyber Resilience“, SECURE. Zugegriffen: 30. September 2025. [Online]. Verfügbar unter: https://secure4sme.eu/

Künstliche Intelligenz in der Produktion

Franz Sümecz
Pilotfabrik, Technische Universität Wien
franz.suemecz@tuwien.ac.at

Barbara Krumay
JKU Business School, Johannes Kepler Universität Linz
barbara.krumay@jku.at

Kurzzusammenfassung

Künstliche Intelligenz spielt in der Produktion eine immer größere Rolle. Das zeigt sich auch in den TBIs, die im Rahmen des EDIH AI5production umgesetzt wurden. Allerdings stehen Unternehmen vor der Herausforderung aus dem immer größer werdenden Angebot die KI-Anwendung auszuwählen, die die aktuelle Problemstellung auch lösen kann. Wissenschaftliche Kategorisierungen können helfen, um diese Auswahl zu vereinfachen. Zusätzlich können Beispiele (Use Cases) genutzt werden, die zusätzlich zu diesen Kategorisierungen als Orientierung für ähnliche Problemstellungen dienen können. Insbesondere Anwendungen, die unterschiedliche Technologien integrieren können aus diesen Beispielen lernen. Im Rahmen dieses Beitrags soll gezeigt werden, wie Kategorien und Use Cases für Unternehmen die Selektion von entsprechenden Ansätzen unterstützt werden kann.

Keywords: Künstliche Intelligenz, Produktion, Anwendungen, Kategorisierung

Abstract

Artificial intelligence is playing an increasingly important role in production. This is also evident in the TBIs that were implemented as part of EDIH AI5production. However, companies are faced with the challenge of selecting the AI application that can solve their current problem from an ever-growing range of options. Scientific categorisations can help to simplify this selection process. In addition, examples (use cases) can be used to provide guidance for similar problems in addition to these categorisations. Applications that can integrate different technologies can learn from these examples in particular. This article aims to show how categories and use cases can support companies in selecting appropriate approaches.

Keywords: Artificial intelligence, production, application, categorization

Einleitung

Die Bedeutung von Systemen, die auf Künstlicher Intelligenz (KI) basieren, wächst ständig weiter. Für die Produktion der Zukunft werden sie unentbehrlich sein. Bereits im Jahr 2018 erwarteten drei von vier Unternehmen, dass KI-Anwendungen das Kerngeschäft innerhalb der nächsten drei Jahre revolutionieren werden [1]. Prognosen gehen davon aus, dass insbesondere Generative KI das globale Bruttoinlandsprodukt um 7 % erhöhen könnte, gleichzeitig aber auch 300 Millionen Arbeitsplätze von „Knowledge Workers“ verloren gehen könnten [2]. Unternehmen erwarten, dass durch den Einsatz von KI Effizienz und Effektivität gesteigert werden kann [3]. Allerdings hat sich gezeigt, dass die Auswahl eines passenden KI-basierten Systems schwierig ist, da die erwünschten Ergebnisse und der langfristige Nutzen schwer einschätzbar sind. Große und disruptive KI-Projekte scheitern häufig, während kleinere, anwendungsorientierte Projekte (als „Low Hanging Fruits“ bezeichnet), die auf bereits existierenden Anwendungen basieren, erfolgversprechender sind [1], [4]. Die Vielzahl der Anwendungsmöglichkeiten - von analytischen Aufgaben über Entscheidungsunterstützung in Echtzeit bis hin zur Automatisierung in der Produktionshalle – führen dazu, dass eine Auswahl über die herkömmliche Make-or-Buy-Entscheidung hinausgeht [5]. Im Vordergrund für diese Auswahl steht meist ein real existierendes Problem, das in der Regel nicht durch eine Technologie, sondern nur im Zusammenspiel unterschiedlicher Technologien (z.B. Internet of Things und KI im Rahmen von Industrie 4.0) gelöst werden kann [5]. Unerheblich ist für diese Entscheidung, ob unter dem Begriff KI Machine Learning (ML), Deep Learning oder Generative KI verstanden werden. Wichtig ist vielmehr zu verstehen, dass

KI-basierte Systeme keinen Selbstzweck haben, sondern nur dann sinnvoll eingesetzt werden, wenn die zugrundeliegende Problemstellung und das Ziel, das mit dem Einsatz von KI-basierten System erreicht werden soll, klar sind. Zu diesen Zielen zählen die Maximierung des Betriebsergebnisses, die Steigerung des Umsatzes, die Senkung der Kosten, die Stärkung der Wettbewerbsfähigkeit, die Verbesserung der Resilienz insbesondere in globalen Wertschöpfungsketten sowie die Stärkung der Nachhaltigkeit (Ressourcenschonung, Kreislaufwirtschaft, Energiewende) [1], [3], [5], [6]. Die Einführung und spätere Nutzung von KI-basierten Systemen – insbesondere für produzierende Unternehmen – basiert unter anderem auch auf einer Möglichkeit, die Systeme und deren Anwendbarkeit in kleinem Rahmen zu testen [1], [7]. Die im EDIH AI5production durchgeführten TBIs sind dafür ein geeigneter Ansatz. Basierend auf den Problemstellungen in den TBIs wurden Kategorien identifiziert, die für andere produzierende Unternehmen als Basis für eine Entscheidung herangezogen werden können.

Künstliche Intelligenz

Die Europäische Union definiert ein KI-System als „ein maschinengestütztes System, das so konzipiert ist, dass es mit unterschiedlichem Grad an Autonomie betrieben werden kann und nach seiner Einführung Anpassungsfähigkeit zeigt, und das für explizite oder implizite Ziele aus den Eingaben, die es erhält, ableitet, wie es Ausgaben wie Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen generieren kann, die physische oder virtuelle Umgebungen beeinflussen können“ [8]. KI imitierte menschliche kognitive Fähigkeiten (Intelligenz), um komplexe Problemstellungen automatisiert zu lösen, die bislang nur durch menschliche Eingriffe (menschliche Intelligenz) im Sinne einer Beurteilung und Entscheidungsfindung zu lösen waren [9]. Erste wissenschaftliche Quellen zum Thema KI stammen bereits aus dem Jahr 1955 [10] - seitdem hat sich das Verständnis nicht grundlegend geändert, die Anwendungsmöglichkeiten aber schon. Die Grundlage für alle KI-basierten Systeme sind Daten, die in strukturierter oder unstrukturierter Form vorliegen können. Wie für alle datenbasierten Anwendungen gilt auch für KI-basierte Systeme, dass die Qualität der Daten die Ergebnisse beeinflussen („Gargabe in – garbage out“). Auf diese Daten werden unterschiedliche Verfahren angewandt (z.B. statistische Verfahren) um eine bestimmte Problemstellung zu lösen. Ein grundlegender Unterschied zu Datenverarbeitung ohne KI ist, dass die KI auf einem kleinen Teil der Daten „trainiert“ wird, um daraus Möglichkeiten zur Interpretation der gesamten Datenbasis zu entwickeln [5]. Bestehende Kategorisierungen von KI-basierten Systemen beziehen sich vor allem auf die Art und Weise, wie Daten genutzt werden. Die Zusammenhänge werden häufig in einem mehrschichtigen Modell dargestellt. KI bildet dabei die äußere Schicht mit Machine Learning (ML) als Teilmenge von KI und Deep Learning (DL) als Teilmenge von ML.

Eine besondere Bedeutung hat die sogenannte generative KI erlangt. Generativ bezieht sich auf die automatisierte Generierung von neuen Texten, Bildern, Videos oder Musik [6]. Neu bedeutet in diesem Zusammenhang, dass das Ergebnis in den zugrundeliegenden Daten in dieser Form nicht vorhanden ist. Viele Anwendungen der generativen KI basieren auf der direkten Interaktion mit Menschen. Diese Interaktion wird meist in Form eines Prompts ermöglicht, das heißt es wird natürlichsprachlich und in Textform das gewünschte Ergebnis vom Menschen in einem textbasierten Interface beschrieben. Daraus wird automatisiert ein Ergebnis generiert, das dieser Beschreibung möglichst entspricht. In einem Dialog (Chat) zwischen Menschen und KI kann das Ergebnis weiter verfeinert werden. Das Ergebnis kann je nach verwendetem Modell spezifisch ein Text (z.B. ChatGPT - GPT-o1) oder ein Bild (Text-zu-Bild-Generatoren wie z.B. Midjourney, DALL-E, Stable Diffusion) sein. Manche KI-basierten Systeme fokussieren auch gezielt auf die Generierung von Programmcode und übernehmen dadurch Aufgaben der Softwareentwicklung (z.B. Cursor). Im Lauf der Zeit haben sich die Angebote weiterentwickelt und wurden verstärkt generell einsetzbar, d.h. sie können verschiedene Ergebnisse liefern (z.B. ChatGPT 5.0, Google Gemini). Auch die Integration von KI-basierten Systemen in Anwendungen entwickelt sich ständig weiter. Dazu gehören in Office-Anwendungen integrierte Systeme (z.B. Microsofts Co-Pilot, KI-gestützte Adobe Acrobat Funktionen zur Funktionen zur Texterstellung und -bearbeitung) aber auch KI als Teil von Steuerungssysteme [5]. Generative KI erleichtert ganz wesentlich den Büroalltag, insbesondere bei der Erstellung von Dokumenten unterschiedlichster Art. Dies ist derzeit die bekannteste Form der Anwendung von KI-Methoden, die vor allem auch im Consumer-Bereich breiten

Einzug gehalten hat. Large Language Models (LLMs) haben sich hier durchgesetzt, also Systeme die auf großen, textbasierten Daten mit Hilfe statistischer Methoden neuen Text erzeugen können.

Kategorisierung von KI-basierten Systemen

Um das Verständnis für KI-basierte Systemen zu erhöhen aber auch um eine strukturierte Auswahl zu ermöglichen, wurden unterschiedliche Kategorisierung vorgenommen. Viele Kategorienschemata beruhen auf darauf, welche Verfahren auf die Daten angewendet werden. Ein Ansatz der im Kontext von produzierenden Unternehmen entstanden ist fokussiert auf Machine Learning (ML) als oberste Ebene und kategorisiert weiter nach Lernparadigmen („learning paradigms“), Lernmodellen („learning models“), Lernaufgaben („learning tasks“) und ordnet diesen Kategorien Anwendungen in produzierenden Unternehmen zu (siehe Abbildung 1) [5].

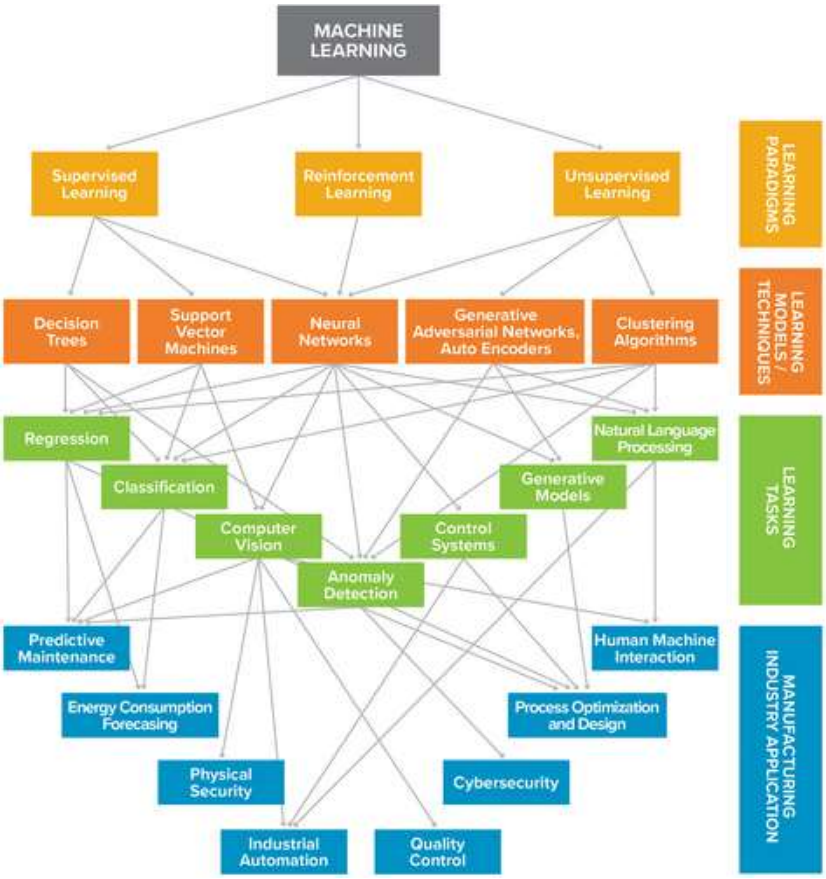

Abbildung 1. Kategorisierung von ML inkl. Anwendungen in der Produktion [4]

Eine andere Kategorisierung, die sich auf Problemstellungen in Unternehmen generell bezieht, unterscheidet lediglich drei Kategorien [1]: (1) Automatisierung von Geschäftsprozessen („automating business processes“), (2) Datenanalyse („gaining insight through data analysis“) und die (3) Interaktion mit Kund:innen und Mitarbeiter:innen („engaging with customers and employees“) [1]. Basierend auf regelbasierte Problemlösungen ohne KI (wie z.B. Geschäftsprozessautomatisierung), wird durch die KI eine andere Form der Problemlösung möglich, die stärker dem menschlichen Vorgehen gleicht. Die Automatisierung von Geschäftsprozessen wird durch KI-Unterstützung in Richtung Robotic Process Automation (RPA) erweitert, die z.B. Verträge lesen und interpretieren kann um Prozessschritte anzustoßen. Ähnliches trifft auf KI-gestützte Datenanalyse zu, die vor allem mit Hilfe von Machine Learning-Ansätzen bestimmte Muster erkennen („Pattern recognition“) und interpretieren kann, um z.B. notwendige Wartungsarbeiten vorherzusagen („predictive maintenance“). Die letzte Kategorie nach beschreibt textbasierte Anwendungen (z.B. Chatbots), die im Kunde:innenservice, für Vorschläge zur Medikation von Patient:innen aber auch im internen Wissensmanagement eingesetzt werden können.

Implementierung von KI-basierten Systemen

Wie bereits erwähnt, ist die Implementierung und langfristige Nutzung von KI-basierten Systemen nicht immer erfolgreich. Es hat sich gezeigt, dass dies vor allem dann gelingt, wenn KI-basierte Systeme integriert in oder zusätzlich zu bereits existierenden Anwendungen implementiert werden [1]. Vier entscheidende Faktoren wurden identifiziert: (1) Technologieverständnis, (2) Projektportfolios, (3) Pilotierung und (4) Skalierung [1]. Technologieverständnis bedeutet, dass im Unternehmen ein gewisses Verständnis für KI-Technologien bestehen muss, insbesondere hinsichtlich deren Stärken und Schwächen. Da im Unterschied zu regelbasierten Systemen keine vollständige Transparenz erreicht werden kann, ist auch Vertrauen ein wichtiges Thema. Darüber hinaus muss ein Projektportfolio vorhanden sein, das die Möglichkeiten von KI-basierten Systemen im Unternehmen abdeckt und anhand von Use-Cases aufzeigt. Eine besondere Rolle kommt der Pilotierung zu, d.h. KI-basierte Systemen müssen in kleinem Rahmen getestet werden, um im Sinne eines Proof-of-concept die Wirksamkeit der Systeme auch – unter Berücksichtigung von Regularien - einschätzen zu können, bevor große Investitionen unternommen werden. Darüber hinaus wird so auch das Vertrauen in das KI-basierte System gestärkt. Wird das KI-basierte System nach der Pilotierung tatsächlich Implementiert, muss darauf geachtet werden, dass es auch skaliert, d.h. dass es unternehmensweit verwendet werden kann auch für zukünftige Veränderungen (z.B. durch mehr Daten) robust und erweiterbar ist [1].

Vorgehen und Ergebnisse

Im Rahmen des EDIH AI5production soll auch gezeigt werden, wie mit Hilfe von Digitalisierung (im weitesten Sinne des Begriffes) eine Effizienzsteigerung in der Produktion erreicht werden kann. D.h. es geht darum zu zeigen, wie die Nutzung digitaler Technologien zur Erreichung der angeführten Unternehmensziele beitragen kann. In drei Jahren wurden über 150 TBIs im Rahmen des EDIH AI5production durchgeführt. Dabei zeigt sich klar, dass KI-Methoden längst ein integraler Bestandteil digitalisierter Produktionsprozesse und Produktionsabläufe geworden sind. 45% der im EDIH AI5production durchgeführten TBIs testen technische Lösungen unter Anwendung von KI (nachfolgend als TBIs mit KI-Anwendung bezeichnet). Basierend auf den existierenden Kategorienschemata [1], [5] wurden diese TBIs analysiert. TBIs, die Charakteristika mehrere Kategorien enthalten, wurden mehrfach zugeordnet. In einer weiterführenden Analyse wurden Kategorien mit Hilfe von ausgewählten TBIs spezifiziert. Eine Übersicht findet sich in Tabelle 1 (Mehrfachzuordnung). Fast die Hälfte der TBIs mit KI-Anwendung nutzen bildbasierter KI, rund 30 % der TBIs mit KI-Anwendung werden für die Analyse von Daten genutzt. Das liegt auch daran, dass in der Produktion eine Vielzahl an Daten aus unterschiedlichen Quellen mit hohem Potential für KI-Anwendungen anfällt. Im Fokus dieser Anwendungen stehen Kostensenkung und Effizienzsteigerung. Rund 14 % der TBIs mit KI-Anwendung setzen generative KI (insbesondere LLMs) ein, um Text zu verstehen, zusammenzufassen und zu generieren (siehe auch [11]).

Tabelle 1. Anzahl der in den TBIs eingesetzten KI-basierten Systeme nach Kategorie und Ebene [1], [5]

Ebenen nach [5]	**Automatisierung von Geschäftsprozessen [1]**	**Datenanalyse [1]**	**Interaktion mit Kund:innen und Mitarbeiter:innen [1]**
Predictive Maintenance	17	22	2
Energy Consumption Forecast	2	2	1
Physical Security	1	2	0
Industrial Automation	27	17	5
Quality Control	5	11	4
Cybersecurity	2	2	0
Process Optimization and Design	7	22	4
Human Machine Interaction	7	11	7

Im Folgenden sollen nun einige konkrete Fallbeispiele veranschaulicht werden, wie in den TBIs des AI5production KI-Methoden für die Produktion umgesetzt und getestet wurden.

Bildbasierte Anwendungen

Die bildbasierten KI bietet ein enorm breites Anwendungsspektrum und dadurch ein großes Potential um Produktionsprozesse und -abläufe effizienter zu gestalten. Dabei geht es einerseits um die Automatisierung und Steuerung von Produktionsprozessen und andererseits um Qualitätssicherung und Qualitätskontrolle. Ein typisches Beispiel aus dem EDIH AI5production steht im Zusammenhang mit dem Refurbishing von gebrauchten Photovoltaik-Modulen (PV-Module). Die zunehmende Anzahl an PV-Modulen bringt auch das Problem der Wiederaufbereitung mit sich. Im Rahmen eines TBI wurde die Möglichzeit zu einer automatisierten Prüfung von PV-Modulen untersucht. Dabei war eine der Problemstellungen das automatisierte Kontaktieren der Stecker des Moduls mittels eines Roboters. Hier kam eine KI basierte Bildverarbeitung zur Anwendung. Die Lage, Position und Typ des Steckers mussten richtig erkannt werden und mit diesen Daten der Roboter gesteuert werden. Ohne KI-Methoden wäre dieser Automatisierungsschritt nicht realisierbar.

KI-gestützte Datenanalyse

Ein wichtiges Zukunftsthema ist die KI-gestützte Datenanalyse, insbesondere um Anomalien in der Produktion (Produktionsmaschinen und Produktionsanlagen) zu erkennen. Ziel ist es, durch eine frühzeitige Anomalieerkennung die Qualität zu sichern und den Ausschuss zu reduzieren. Unter dem Begriff „Predictive Maintenance“ werden diese Ansätze zusammengefasst. Diese erlauben basierend auf Daten notwendige Wartungen rechtzeitig und kontrolliert durchzuführen zu können, um ungeplante Stillstände zu verhindern. Insbesondere im Zusammenhang mit Spanabhebenden Maschinen (um Späne vom Werkstück abzuheben) kommt es immer wieder zu Qualitätsschwankungen in der Feinbearbeitung und damit zu Ausschuss. In einem TBI wurden Sensordaten der Bearbeitungsmaschine in Echtzeit abgegriffen und mittels KI-basierten Ansätzen für den Domänenexperten aufbereitet. Dadurch konnte ein Verbesserungspotential für den Bearbeitungsprozess ermittelt werden und der Ausschuss reduziert werden (Abbildung 2).

Abbildung 2. Aufspannung für eine spanabhebende Bearbeitungsmaschine

Generative KI

Die bereits angesprochene populäre Generative KI biete abseits der Erleichterung des Büroalltags auch in der Produktion viele Anwendungsmöglichkeiten. Zwei konkrete Beispiele sollen vorgestellt werden.

Für produzierende Unternehmen ist der Einkauf von Rohstoffen ein wichtiger Kosten- und Qualitätsfaktor. Bei Rohstoffen, die zu Tageskursen am Weltmarkt gehandelt werden, ist es für das Betriebsergebnis von großer Bedeutung im richtigen Augenblick zum richtigen Preis die richte Menge einzukaufen. Weltweite Ereignisse (Wetterkatastrophen, politische Spannungen, Kriegsereignisse, Pandemien, Nachfrageprognosen, etc.) haben Einfluss auf diese Weltmarktpreise. KI-Systeme bieten hier die Möglichkeit, die komplexen Zusammenhänge zu analysieren um Handlungsempfehlungen abzuleiten In diesem Zusammenhang wurde ein TBI durchgeführt um das Potential von KI-basierten Stahlmarktanalysen für den weltweiten Stahleinkauf zu untersuchen.

Ein Problem mit dem viele produzierende Unternehmen konfrontiert sind ist, dass die zur Verfügung stehenden Maschinendaten nur unzureichend oder gar nicht genutzt werden können, weil es an Schnittstellen und Data-Scientists fehlt. Als fiktives Beispiel: Ein Maschinenführer weiß, dass steigende Temperatur und gleichzeitig abfallender Druck in einem Hydraulikaggregat ein Problem darstellen. Dies wäre auch in den Maschinendaten ersichtlich, es kann aber kein automatisierter Alarm-Trigger gesetzt werden, ohne geeignete Schnittstelle, Data-Analyst und/oder Programmierer. An diesem Punkt setzt ein TBI an, in dem ein Tool konzeptioniert und entwickelt werden soll, welches mit KI-basiert eine individuelle Problembeschreibung in natürlicher Sprache verstehen und diesen Input in die notwendigen Scripte (R, Phyton) übersetzen kann, um automatisiert Auswertungen oder Alarme zu erstellen. Damit wird das Wissen der Mitarbeiter:innen optimal gespeichert und genützt. Im Rahmen eines TBIs wurde ein KI-Prototyp entwickelt, der natürlichsprachlichen Input verarbeiten kann, die notwendigen Daten geeignet aggregiert und verarbeitet und schlussendlich Reports erstellen oder Alarm-Trigger setzt. Dieser KI-Assistent erleichtert die Arbeit des Maschinenbedieners und ermöglicht, dass dessen Know-How im System erfasst wird. Am Ende des TBI steht ein Prototyp, anhand dessen abgeschätzt werden kann, ob eine Weiterentwicklung für das Unternehmen zielführend ist.

Assistenzsysteme

Eine besondere Rolle nehmen Assistenzsysteme ein, die KI in Kombination mit anderen Technologien (z.B. Augmented Reality) nutzen, um den Menschen bei wiederkehrenden komplexen oder monotonen Arbeitsschritten zu unterstützen. Dadurch kann die Belastung des Menschen gesenkt und die Fehlerquote reduziert werden. Mit Hilfe von Augmented Reality (AR) und Projektionssystemen können Markierungen angezeigt werden, die dem Menschen Informationen über den nächsten Arbeitsschritt bieten. Beispielsweise die farbliche Lichtmarkierung von Teilen nach einem Fertigungsprozess auf Basis von Produktionsplänen. Durch die Lichtprojektion dem Menschen Informationen über den Teil für den nächsten Arbeitsschritt vermittelt (siehe Abbildung 3).

Abbildung 3. Assistenzsysteme in der Produktion – AR-Anwendung

Ein weiteres Beispiel für Assistenzsysteme betrifft die Arbeitsvorbereitung. Die Aufgabenstellung hierbei ist die korrekte Aufspannung von Bauteilen nach Vorgaben für die spanabhebende Bearbeitung. Gemeinsam mit einem größeren Unternehmen wurden die Möglichkeiten für den Einsatz eines Assistenzsystems untersucht, mit dessen Hilfe die korrekten Spannpunkte auf die Spannvorrichtung projiziert werden. Der Vorgang in der Arbeitsvorbereitung wird dadurch um 50% verkürzt. In diesem Projekt sind die KI-basierten Systeme integriert und kommen mehrfach zum Einsatz, zum Beispiel für die Erkennung der einzelnen Strukturelemente und die korrekte Projektion.

Ein anderer Anwendungsbereich ist die Instandhaltung und Wartung von Maschinen und Anlagen oder auch von Flugzeugen. Im Zusammenhang mit der Wartung von Flugzeugen wurde ein TBI durchgeführt, bei dem ein Konzept für den Einsatz eines AR-basierten Systems entwickelt wurde. Als Technologie kann hier die Microsoft HoloLens 2 unterstützt durch KI verwendet werden. Durch die Integration eines unterstützenden Deep Learning Modells ist das System ist in der Lage, eigenständig Bereiche zu identifizieren und darauf aufbauend eine Aussage über den Ist-Zustand des Systems zu treffen. Eine effektive Integration von AR und KI in die Instandhaltungsprozesse kann einen erheblichen Einfluss auf das Zuverlässigkeitsmanagement (i.e., Erhöhung von Effizienz oder Genauigkeit) haben.

Zusammenfassung

Fast die Hälfte der im EDIH AI5production umgesetzten TBIs beziehen auf unterschiedliche Weise KI-basierte Methoden mit ein. Der Fokus liegt dabei nicht auf der KI-Technologie, sondern viel entscheidender ist die Anwendung in der Produktion um eine konkrete Problemstellung automatisiert zu lösen. Neben einem generellen Einsatz von Generativer KI konnten drei Anwendungsgebiete identifiziert werden: Bildbasierte KI, Assistenzsysteme und KI-gestützte Datenanalyse. Dies entspricht weitgehend auch dem Stand der Forschung [1]. Es hat sich in den TBIs auch gezeigt, dass dadurch für die Unternehmen komplexe Problemstellungen, die bislang eine menschliche Intervention erforderlich machten, zu automatisieren. Gezeigt wurde auch, dass Effizienz und Effektivität gesteigert werden können, zum Beispiel in Form von Erhöhung der Geschwindigkeit und Verkürzung der Taktzeiten, Senkung der Kosten und Reduktion des Ausschusses und der Fehlerrate. Die Anwendung von KI-Methoden erfolgt in vielen TBIs in integrierter Form und ist daher nicht offensichtlich erkennbar. Dennoch ermöglichen sie völlig neue Stufen der Automatisierung und Assistenz.

Für Unternehmen generell, aber insbesondere produzierende Unternehmen, lassen sich aus den Ergebnissen und den wissenschaftlichen Erkenntnissen unterschiedliche Handlungsempfehlungen ableiten. Durch eine Kategorisierung der Problemstellung kann die Auswahl von KI-Systemen für die Produktion treffsicherer aber auch schneller erfolgen. Wie bereits in unterschiedlichen Studien identifiziert [1], [5] können KI-basierte Systeme dann erfolgreich eingeführt werden und langfristig betrieben werden, wenn gewisse Faktoren erfüllt sind. Dazu gehören einerseits Technologieverständnis (z.B. durch Schulungen) aber auch Pilotieren (z.B. durch TBIs). Im Fokus sollten Projekte stehen, die auf existierenden Anwendungen basieren und die KI zur Problemlösung als Erweiterung gesehen wird. Dies führt zu einer höheren Akzeptanz des Systems, da die KI nicht als Bedrohung, sondern im besten Fall als Arbeitserleichterung empfunden wird. Besonders deutlich zeigt sich das in den Assistenzsystemen, die für die Mitarbeiter:innen eine Verbesserung gebracht haben. Ein weitere wichtiger Aspekt ist die Berücksichtigung der Datenlage, da KI-basierte Systeme nur so gut sind wie die dahinterliegenden Daten. Daher müssen vorab schon Produktionsdaten gesammelt werden, damit die Systeme entsprechend trainiert werden können. Ein Re-Training ist in vielen Fällen von Vorteil. Die in den TBIs eingesetzte Generative KI, insbesondere LLMs, können gerade in produzierenden Unternehmen auch in Form von Narrow Language Models, also mit kleineren, unternehmensinternen Datensätzen umgesetzt werden, um etwaige Verzerrungen (Biases) in frei verfügbaren oder zugekauften Daten zu vermeiden. Diese Aspekte sollten in die Entscheidungsfindung miteinbezogen werden, wenn KI-basierte Systeme in produzierenden Unternehmen eingesetzt werden sollen.

Literatur

[1] T. H. Davenport und R. Ronanki, „Artificial intelligence for the real world“, *Harvard business review*, Bd. 96, Nr. 1, S. 108–116, 2018.
[2] „Generative AI could raise global GDP by 7%“. Zugegriffen: 19. September 2025. [Online]. Verfügbar unter: https://www.goldmansachs.com/insights/articles/generative-ai-could-raise-global-gdp-by-7-percent
[3] F. Uche Ojika, W. Oseremen Owobu, O. Anthony Abieba, O. Janet Esan, B. Chibunna Ubamadu, und A. Ifesinachi Daraojimba, „The Role of Artificial Intelligence in Business Process Automation: A Model for Reducing Operational Costs and Enhancing Efficiency“, *IJAMRS*, Bd. 4, Nr. 6, S. 1449–1462, Dez. 2024, doi: 10.62225/2583049X.2024.4.6.4046.
[4] I. M. Enholm, E. Papagiannidis, P. Mikalef, und J. Krogstie, „Artificial Intelligence and Business Value: a Literature Review“, *Inf Syst Front*, Bd. 24, Nr. 5, S. 1709–1734, Okt. 2022, doi: 10.1007/s10796-021-10186-w.
[5] S. J. Plathottam, A. Rzonca, R. Lakhnori, und C. O. Iloeje, „A review of artificial intelligence applications in manufacturing operations“, *Journal of Advanced Manufacturing and Processing*, Bd. 5, Nr. 3, S. e10159, 2023, doi: 10.1002/amp2.10159.
[6] S. Feuerriegel, J. Hartmann, C. Janiesch, und P. Zschech, „Generative AI“, *Bus Inf Syst Eng*, Bd. 66, Nr. 1, S. 111–126, Feb. 2024, doi: 10.1007/s12599-023-00834-7.
[7] N.-A. Perifanis und F. Kitsios, „Investigating the influence of artificial intelligence on business value in the digital era of strategy: A literature review“, *Information*, Bd. 14, Nr. 2, S. 85, 2023.
[8] „Article 3: Definitions | EU Artificial Intelligence Act“. Zugegriffen: 19. September 2025. [Online]. Verfügbar unter: https://artificialintelligenceact.eu/article/3/
[9] „Künstliche Intelligenz (KI) und maschinelles Lernen - Fraunhofer IKS“, Fraunhofer-Institut für Kognitive Systeme IKS. Zugegriffen: 19. September 2025. [Online]. Verfügbar unter: https://www.iks.fraunhofer.de/de/themen/kuenstliche-intelligenz.html
[10] J. McCarthy, M. L. Minsky, N. Rochester, und C. E. Shannon, „A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955“, *AI Magazine*, Bd. 27, Nr. 4, S. 12–12, Dez. 2006, doi: 10.1609/aimag.v27i4.1904.
[11] F. Sümecz, „AI5production - Der Weg zur digitalen Transformation in der Produktion“. TU Wien, 2025.

Digitale Datendurchgängigkeit für die Antriebstechnik in der Gebäudetechnik

Gerhard Zucker
AIT Austrian Institute of Technology
gerhard.zucker@ait.ac.at

Tanja Ivanovic
AIT Austrian Institute of Technology
tanja.ivanovic@ait.ac.at

Kurzzusammenfassung

Das Projekt DRIVE_THRU zielte auf die Schaffung einer durchgängigen digitalen Datenbasis für Komponenten der technischen Gebäudeausrüstung (TGA), insbesondere Antriebstechnik, ab. Auf Basis offener Standards wie IFC (Industry Foundation Classes), dem offenen Format für Building Information Modeling (BIM), wurden Methoden entwickelt, um Herstellerdaten nahtlos in Planungs-, Ausführungs- und Betriebsprozesse zu integrieren. Die Ergebnisse umfassen eine BIM-konforme Revit-Demofamilie mit Parametern für Lebenszyklusanalyse (LCA) und Kreislaufwirtschaft sowie einen prototypischen Workflow zur Datenweitergabe über den gesamten Lebenszyklus. Die dadurch gewonnene Datenkonsistenz und Transparenz unterstützt fundierte Entscheidungen in Planung, Ausführung und Betrieb und leistet einen Beitrag zu nachhaltigem Bauen.

Keywords: Building Information Modeling (BIM), Gebäudetechnik, Antriebstechnik, Kreislaufwirtschaft, Life Cycle Assessment (LCA)

Abstract

The DRIVE_THRU project aimed to create a comprehensive digital database for technical building equipment (TBE) components, particularly drive technology. Based on open standards such as IFC, methods were developed to seamlessly integrate manufacturer data into planning, execution and operational processes. The results include a BIM-compliant Revit demo family with parameters for life cycle analysis (LCA) and circular economy, as well as a prototypical workflow for data transfer throughout the entire life cycle. The resulting data consistency and transparency supports informed decisions in planning, execution and operation and contributes to sustainable construction.

Keywords: Building Information Modeling (BIM), Life Cycle Assessment (LCA)

Einleitung

In der Bauwirtschaft werden große Fortschritte hinsichtlich Digitalisierung und Nachhaltigkeit angestrebt. Technische Gebäudeausrüstung (TGA) spielt eine zentrale Rolle für die Energieeffizienz und Betriebsqualität von Gebäuden, doch gerade in diesem Bereich fehlt es häufig an einer durchgängigen, strukturierten Datenverfügbarkeit über den gesamten Lebenszyklus hinweg. Hersteller stellen zwar zunehmend digitale Produktinformationen zur Verfügung, diese sind jedoch oft nicht mit Planungswerkzeugen kompatibel oder für spätere Nutzungsphasen nutzbar. Die Planung erfolgt vielfach auf Basis unvollständiger oder redundanter Informationen, was zu Ineffizienzen und Medienbrüchen führt. Das Projekt DRIVE_THRU adressiert diese Problematik am Beispiel von Antriebstechnik-Komponenten und entwickelt ein prototypisches Datenmodell, das Informationen von der Produktion über die Planung und Ausschreibung bis hin zum Betrieb durchgängig verfügbar macht, mit besonderem Augenmerk auf der Integration von Nachhaltigkeitsaspekten wie Lebenszyklusdaten und Kreislaufwirtschaft.

Ausgangssituation und Zielsetzung

Die Projektpartner identifizierten eine zentrale Herausforderung in der mangelnden Interoperabilität und fehlenden semantischen Konsistenz technischer Produktdaten in BIM-Prozessen. Antriebskomponenten

müssen in frühen Phasen der Gebäudeplanung als generische Bauteile modelliert werden, wobei wesentliche technische, ökologische und betriebliche Informationen fehlen oder erst später ergänzt werden. Diese Diskontinuität führt zu Informationsverlusten und erhöhtem Koordinationsaufwand. Das Ziel des Projekts war es, eine einheitliche, offene Datenstruktur zu entwickeln, die sowohl die generischen als auch die produktspezifischen Informationen abbildet und in gängigen Softwareumgebungen nutzbar macht. Durch die Nutzung offener Standards wie IFC [1] sollte die Grundlage für eine breitere Implementierung im Sektor geschaffen werden. Ergänzend dazu war es Ziel, ausgewählte Umweltkennzahlen in Form von LCA-Parametern [2] in das Modell zu integrieren, um ökologische Aspekte systematisch in die Planung einzubeziehen.

Methodik

Im Zentrum der methodischen Umsetzung stand die systematische Analyse von Informationsanforderungen entlang des Lebenszyklus technischer Komponenten. Dafür wurden Interviews mit Expertinnen und Experten aus den Bereichen Planung, Ausführung und Gebäudebetrieb geführt, um die spezifischen Informationsbedarfe in jeder Phase zu erfassen. Parallel wurden Umweltproduktdeklarationen (EPDs) analysiert, um belastbare Umweltkennzahlen zu extrahieren. Aus diesen Erkenntnissen wurde ein konsistenter Parametersatz abgeleitet, der sowohl betriebliche als auch ökologische Informationen umfasst. Diese Parameter wurden in eine prototypische Revit-Familie integriert, die als Demonstrator dient und zeigt, wie eine strukturierte Datenmodellierung für Antriebstechnik aussehen kann. Die technische Umsetzung erfolgte dabei unter Berücksichtigung des IFC-Schemas gemäß ISO 16739, um maximale Interoperabilität zu gewährleisten. Neben der Modellierung der Parameter wurde ein exemplarischer Workflow entwickelt, der die Weitergabe und Nutzung dieser Daten in verschiedenen Softwareumgebungen und Anwendungsphasen illustriert.

Ergebnisse und Demonstrator

Ein zentrales Ergebnis des Projekts ist die Entwicklung einer standardisierten, BIM-kompatiblen Revit-Familie für ausgewählte Komponenten unseres Projektpartners Belimo Automation. Diese Familie enthält eine Vielzahl von Parametern, die sowohl technische Merkmale wie Energieverbrauch, Anschlusswerte oder Medienkompatibilität als auch ökologische Kennzahlen wie Global Warming Potential (GWP), Rezyklierbarkeit oder Lebensdauer abbilden. Zusätzlich wurden betriebsrelevante Daten wie Seriennummern oder Wartungszyklen berücksichtigt, um die Nutzbarkeit im Facility Management zu erhöhen. Der Demonstrator zeigt nicht nur die strukturierte Parametrierung, sondern auch die Übergabepunkte zwischen den einzelnen Lebenszyklusphasen, wie z. B. die Übertragung generischer Ausschreibungsdaten in herstellerspezifische Planungsmodelle. Der prototypische Workflow illustriert, wie diese Daten automatisiert übernommen, weiterverarbeitet und schließlich für Betriebszwecke verwendet werden können. Damit wird eine digitale Kette etabliert, die von der ersten Planungsidee bis zum Gebäudebetrieb reicht. Abbildung 1 zeigt das Belimo Energy Valve, das als Referenz für den Demonstrator herangezogen wurde.

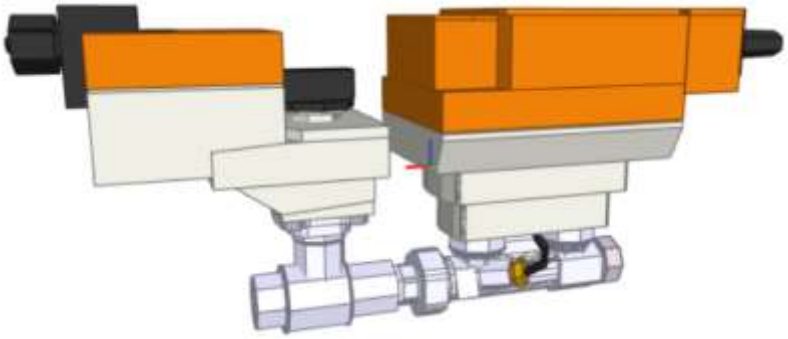

Abbildung 1. Belimo Energy Valve

Ergänzende Analysen: Lebenszyklus, Szenarien und Materialwirkungen

Ein zentrales Element im Projekt DRIVE_THRU war die vertiefte Betrachtung der ökologischen Auswirkungen technischer Komponenten über deren gesamten Lebenszyklus hinweg. Im Fokus stand die Lebenszyklusanalyse (LCA) eines exemplarischen Produkts – konkret des Energy Valve EVO15R2+BAC. Ziel war es, unterschiedliche Lebenszyklusszenarien hinsichtlich ihres CO_2-Fußabdrucks und der Ressourcennutzung zu bewerten. Hierzu wurden zwei Varianten einander gegenübergestellt: das „cradle to grave“-Szenario, das den klassischen linearen Lebensweg von der Rohstoffgewinnung bis zur Entsorgung abbildet, sowie das „cradle to cradle“-Szenario, das auf Materialkreisläufe und Wiederverwendung fokussiert.

Die Berechnungen basierten auf anerkannten Normen wie EN 15804 [3] und ISO 14025 [4] und berücksichtigten sowohl den verkörperten Kohlenstoff (embodied carbon) als auch die Emissionen während der Nutzungsphase (operational carbon). Bei einer angenommenen Gebäudelebensdauer von 100 Jahren und einer Produktlebensdauer von 10 Jahren musste das betrachtete Bauteil insgesamt neunmal ersetzt werden. Im „cradle to grave“-Modell summierte sich der Materialbedarf vollständig aus Primärmaterial, während im „cradle to cradle“-Ansatz ein Großteil durch Sekundärmaterial gedeckt werden konnte. Dies führte zu einer signifikanten Reduktion des ökologischen Fußabdrucks. So zeigte sich, dass 78 % der Produktmasse recycelt und die verbleibenden 22 % energetisch verwertet werden konnten – ohne dass Material deponiert werden musste.

Diese quantitativen Ergebnisse wurden genutzt, um ein konkretes Verständnis für die Relevanz ökologischer Parameter in der digitalen Modellierung zu entwickeln. Die zugrunde liegenden Materialdaten und Nachhaltigkeitskennzahlen wurden direkt in die IFC-Datenstruktur überführt und als erweiterte Parameter in die prototypische Revit-Familie eingebettet. Damit wurde nicht nur ein Beleg für die Umweltleistung des Produkts erbracht, sondern auch die technische Voraussetzung geschaffen, um LCA-Kennzahlen automatisiert in digitale Planungs- und Entscheidungsprozesse zu integrieren.

Innovation und Nutzen für die Industrie

Das Projekt DRIVE_THRU leistet einen wesentlichen Beitrag zur Digitalisierung der TGA-Branche, insbesondere im Hinblick auf die Datenverfügbarkeit und Nachhaltigkeitsintegration. Die Kombination aus offenen Standards, Herstellerdaten und Umweltparametern stellt ein Novum dar, da bisherige Datenmodelle entweder technisch oder ökologisch fokussiert waren, jedoch selten beides integrieren konnten. Für Hersteller bedeutet die strukturierte Aufbereitung ihrer Produktdaten einen Wettbewerbsvorteil, da sie so in digitale Planungsprozesse integriert und von Planern bevorzugt berücksichtigt werden können. Planer profitieren von der verbesserten Datenverfügbarkeit und können frühzeitig fundierte Entscheidungen hinsichtlich Nachhaltigkeit und Betrieb treffen. Betreiber schließlich erhalten vollständige, maschinenlesbare Produktinformationen, was die Grundlage für datenbasiertes Facility Management bildet. Die im Projekt gewonnenen Erkenntnisse fließen direkt in ein geplantes Nachfolgeprojekt ein, das unter dem Titel "TGA Connect" die Standardisierung dieser Parameter auf nationaler Ebene vorantreibt.

Zusammenfassung

DRIVE_THRU demonstriert, wie durch konsequente Digitalisierung und Nutzung offener Standards ein durchgängiger Informationsfluss für Komponenten der TGA geschaffen werden kann. Die strukturierte Modellierung technischer und ökologischer Parameter in einer Revit-Familie sowie die Entwicklung eines interoperablen Workflows zeigen praxisnah auf, wie Herstellerdaten für Planung und Betrieb nutzbar gemacht werden können. Durch die Verbindung von Datenmodellierung, Nachhaltigkeitsbetrachtung und Standardisierung wird ein relevanter Beitrag zur digitalen Transformation und ökologischen Optimierung des Gebäudesektors geleistet.

Danksagung

Wir bedanken uns bei unserem Projektpartner Belimo Automation Handelsgesellschaft m.b.H. für die ausgezeichnete Zusammenarbeit im Projekt und die Unterstützung bei der Erarbeitung der Projektergebnisse.

Literatur

[1] ISO International Organization for Standardization, *EN ISO 16739:2021 - Industry Foundation Classes (IFC) for data sharing in the construction and facility management industries*. 2021. Zugegriffen: 10. April 2020. [Online]. Verfügbar unter: https://www.austrian-standards.at/de/shop/din-en-iso-16739-1-2021-11~p2602356

[2] *BS EN 15978:2011 Sustainability of construction works. Assessment of environmental performance of buildings. Calculation method*. Zugegriffen: 23. Juli 2025. [Online]. Verfügbar unter: https://www.en-standard.eu/bs-en-15978-2011-sustainability-of-construction-works-assessment-of-environmental-performance-of-buildings-calculation-method

[3] *ÖNORM EN 15804: Nachhaltigkeit von Bauwerken - Umweltproduktdeklarationen - Grundregeln für die Produktkategorie Bauprodukt*. Zugegriffen: 23. Juli 2025. [Online]. Verfügbar unter: https://www.austrian-standards.at/de/shop/onorm-en-15804-2022-02-15~p2614831

[4] A. Standards, *ISO 14025: Umweltkennzeichnungen und -deklarationen - Typ III Umweltdeklarationen - Grundsätze und Verfahren*. Zugegriffen: 23. Juli 2025. [Online]. Verfügbar unter: https://www.austrian-standards.at/de/shop/onorm-en-iso-14025-2010-07-01~p1810683

Effizienzsteigerungen bei WIG- und Plasma-Schweißinvertern

Conrad Gstöttner
Austrian Institute of Technology GmbH
Conrad.Gstöttner@ait.ac.at

Andreas Pfeffer
Austrian Institute of Technology GmbH
Andreas.Pfeffer@ait.ac.at

Peter Jonke
Austrian Institute of Technology GmbH
Peter.Jonke@ait.ac.at

Katharina Machtinger
Austrian Institute of Technology GmbH
Katharina.Machtinger@ait.ac.at

Kurzzusammenfassung

Aufgrund des hohen Leistungsbedarfs von Schweißinvertern ist die Energieeffizienz ein wichtiges Qualitätsmerkmal dieser Geräte, weswegen Energieeffizienzmaßnahmen im Fokus dieses TBI standen. Im Rahmen des Projektes wurde moderne Laborinfrastruktur des Austrian Institute of Technology (AIT) verwendet, um Messungen an einem Schweißinverter durchzuführen, der dem aktuellen Industriestandard entspricht. Auf Basis dieser Messungen, sowie Literatur- und Patentrecherchen wurden Verbesserungspotentiale aufgezeigt und Möglichkeiten diskutiert diese auszuschöpfen. Aufgrund des Reifegrades der aktuellen Technologie ist der größte Hebel durch einen Umstieg auf moderne Wide-Bandgap-Halbleiter in Kombination mit einem entsprechend angepasstem Schaltungsdesign und innovativen Regelungskonzepten gegeben. Allerdings ist die bisher geringe Verbreitung dieser Technologien in diesem Anwendungsgebiet ein großes Hemmnis für die Industrie. Daher wurden im Rahmen dieses TBI moderne digitale Möglichkeiten wie Hardware-in-the-Loop-Tests (HIL-Tests) aufgezeigt, welche den Entwicklungsprozess beschleunigen können.

Keywords: Schweißinverter, Leistungselektronik, Wide-Bandgap-Halbleiter, Effizienzsteigerung

Abstract

The high electrical energy demand of welding inverters makes their energy efficiency a key quality feature, which is why measures to increase the energy efficiency were of primary interest for this TBI. Within the project, modern laboratory infrastructure of the Austrian Institute of Technology (AIT) was used to take measurements on an industrial-standard welding inverter. Based on these measurements, as well as literature and patent reviews, potentials for raising the energy efficiency were highlighted and possibilities for exploiting them were discussed. Due to the maturity of the current welding inverter technology, the greatest potential lies in transitioning to modern wide-bandgap semiconductors with accordingly adapted circuit designs and innovative control concepts. However, the limited adoption of these technologies in the field of welding inverters poses a barrier for the industry due to prolonged development cycles. Therefore, within the scope of this TBI modern digital methods such as Hardware-in-the-Loop (HIL) testing were demonstrated to accelerate the development process.

Keywords: welding inverters, power electronics, wide-bandgap-semiconductors, increase in energy efficiency

Einleitung

Nach der EU-Richtlinie 2019/1784 zählen Schweißinverter mit einem für 2030 prognostizierten Endenergieverbrauch von 6 TWh, das entspricht 2,4 Mio t CO2-Äquivalenten, zu den energieverbrauchsrelevanten Produkten [1]. Die Richtlinie schreibt unter anderem Mindestanforderungen an die Energieeffizienz von Schweißinvertern vor, beispielsweise 85% Effizienz und maximal 50W Leistungsaufnahme im Leerlauf für Geräte mit dreiphasiger Stromversorgung und DC-Ausgang. Mithilfe moderner Wide-Bandgap-Halbleiter wie beispielsweise Siliziumkarbid (SiC)-Halbleiter, die in der Elektromobilität bereits Stand der Technik sind und dort Effizienzen von ca. 98 –

99% erzielen (siehe zum Beispiel [2], [3]), können diese Mindestanforderungen voraussichtlich noch deutlich übertroffen werden. Ein weiterer Vorteil verbesserter Effizienz ist die Reduktion der Abwärme, wodurch die Kühlung des Geräts optimiert und im Idealfall passiv ausgeführt werden kann. Damit kann die Gesamteffizienz weiter gesteigert werden und die Leerlaufleistungsaufnahme reduziert werden. Eine geringere Abwärme ermöglicht auch kompaktere Bauweisen und folglich kleinere Gehäuse, wodurch der Ressourcenbedarf für diese aber auch den Transport sinkt. Insbesondere durch die Langlebigkeit der Schweißinverter von typischerweise 10 Jahren im rauen industriellen Umfeld gilt es die Geräte so früh wie möglich auf höchste Energieeffizienz und kompakte Bauweise zu entwickeln.

In der industriellen Praxis sind derzeit häufig Geräte im Einsatz, die überwiegend mit Standardbaugruppen in Fernost produziert werden. Diese können zwar die aktuellen von der EU geforderten Mindestanforderungen erfüllen, sorgen aber nur für eine geringe Wertschöpfung in Europa und schöpfen das Potential, vor allem betreffend Energieeffizienz, aber auch in Hinblick auf Netzrückwirkungen nicht aus. In Bezug auf den Wirkungsgrad stellt der Umstieg auf moderne Halbleitertechnologien, wie beispielsweise der Einsatz von SiC-Halbleitern, die einzige Möglichkeit dar, um substanzielle Verbesserungen zu erzielen. Das erfordert jedoch mehr als nur den Austausch der Halbleiterschaltelemente, es ist ein neues Design notwendig, um das Potential vollständig auszuschöpfen [4]. Dies stellt insbesondere ein Hemmnis für die Industrie dar, da in diesem Anwendungsgebiet diese Technologien bislang kaum im Einsatz waren. Bezüglich der Netzrückwirkungen sind insbesondere moderne Regelungsverfahren bzw. hochpräzise Ansteuerungen der einzelnen Schaltelemente vielversprechende Optionen auch um gezielt die Präsenz relevanter Oberwellen zu reduzieren.

Problemlösungsweg / Vorgehen

Im Rahmen des TBI wurde moderne Laborinfrastruktur des AIT verwendet, um Messungen an einem Schweißinverter durchzuführen, der dem aktuellen Industriestandard entspricht. Der für die Messungen genutzte Schweißinverter ist sowohl für Wolfram-Inertgas-(WIG-) als auch für Plasmaschweißen geeignet. Gemessen wurden die Effizienz und Netzrückwirkungen an verschiedenen Arbeitspunkten für das WIG- und Plasmaschweißen. Die Messungen, in Kombination mit Literatur- und Patentrecherchen, dienten dazu, Verbesserungspotenziale sowie Möglichkeiten zu deren Ausschöpfung aufzuzeigen. Exemplarisch ist in folgender Abbildung 2 die Effizienz sowie die eingangsseitig gemessene Wirk- und Scheinleistung für die vermessenen Plasmaarbeitspunkte dargestellt.

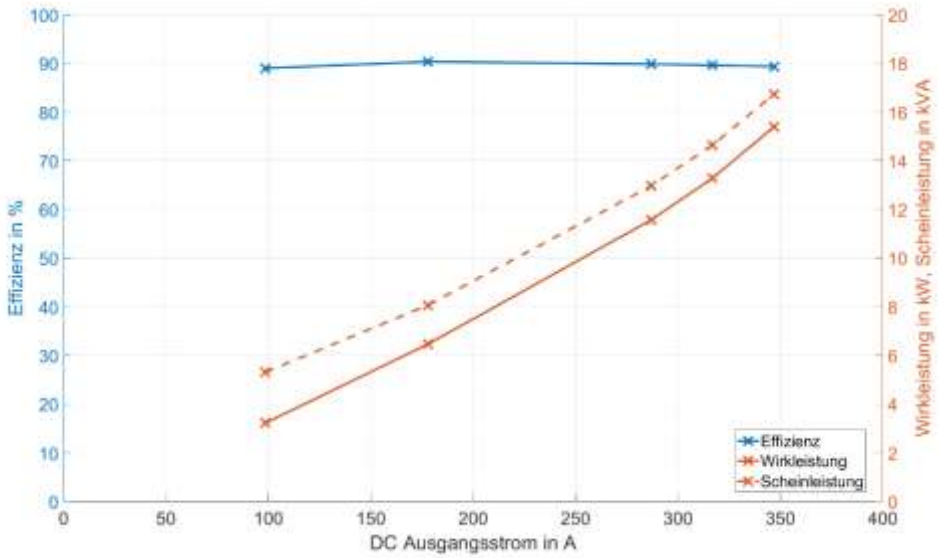

Abbildung 2. Gemessene Effizienz, Wirkleistung und Scheinleistung in verschiedenen Arbeitspunkten für das Plasmaschweißen.

In allen dargestellten Arbeitspunkten für das Plasmaschweißen liegt die Effizienz bei ca. 90%. Das Verhältnis von Wirk- zu Scheinleistung verbessert sich von ca. 0,61 im Arbeitspunkt nahe 100A Ausgangsstrom zu ca. 0,92 im Arbeitspunkt nahe 350A Ausgangsstrom. Die Messung der Netzrückwirkungen ergab zudem, dass die eingangsseitigen Stromverläufe in allen Arbeitspunkten von einem rein sinusförmigen Verlauf abweichen. Insbesondere treten 5. und 7. Harmonische, sowie vereinzelt höhere ungerade Harmonische mit signifikanten Amplituden auf. Dem Hersteller des Schweißinverters wurden die vollumfänglichen Messergebnisse zur Verfügung gestellt.

Die Messergebnisse aus Abbildung 2 verdeutlichen, dass mit den heute in Schweißinvertern verbauten Technologien wie Si-MOSFET bzw. Si-IGBT- Schaltelementen Effizienten im Bereich von 90% erreicht werden. Im Vergleich zu Anwendungen wie der Elektromobilität, in denen moderne Wide-Bandgap-Halbleiterelemente zum Einsatz kommen und damit Wirkungsgrade von 98-99% erreicht werden, zeigt sich jedoch ein deutliches Verbesserungspotential. Die Vorteile von SiC-MOSFETs gegenüber Si-IGBTs in Invertern im Bereich der Elektromobilität wurden Beispielsweise in den Arbeiten [2], [5], [6] und [7] untersucht. Im Bereich von Schweißinvertern sind voraussichtlich Effizienzsteigerungen in einer ähnlichen Größenordnung erzielbar. Aktuelle Forschung zum Einsatz von SiC-Halbleitern in Schweißinvertern kann beispielsweise in [8] und darin zitierten Referenzen gefunden werden.

Hinsichtlich Leistungsfaktor weisen die Messergebnisse insbesondere im Teillastbereich auf signifikantes Verbesserungspotential hin. Eine Reduktion der Oberschwingungen der eingangsseitig gemessenen Stromverläufe wäre in allen Arbeitspunkten erstrebenswert, um auch auf zukünftige gesetzliche Richtlinien zu Netzrückwirkungen vorbereitet zu sein. Zur Verbesserung der Netzrückwirkungen über den gesamten Lastbereich hinweg könnte das im vermessenen Inverter eingesetzte passive Eingangsfilter durch eine Baugruppe zur aktiven Leistungsfaktorkorrektur (englisch: Power Factor Correction, PFC) ersetzt werden. Eine ausführliche Analyse verschiedener Schaltungen zur aktiven Leistungsfaktorkorrektur kann zum Beispiel in [9], [10] gefunden werden.

Ergebnis

Die im vorigen Abschnitt auszugsweise präsentierten Mess- und Rechercheergebnisse, sowie die daraus resultierenden Verbesserungspotentiale und Möglichkeiten selbige auszuschöpfen, wurden mit dem Projektpartner aufbereitet, präsentiert und diskutiert. Die Entwicklung einer neuen Generation von Schweißinvertern, die dem aktuellen Stand hinsichtlich Energieeffizienz, Netzrückwirkungen und Kompaktheit überlegen ist, ist für den Projektpartner, bzw. die Schweißindustrie insgesamt ein erstrebenswertes Ziel. Aufgrund des Reifegrades der aktuell genutzten Technologie ist dafür ein Umstieg auf moderne Wide-Bandgap-Halbleiter mit entsprechend angepasstem Schaltungsdesign und innovativen Regelungskonzepten unabdingbar. Für den Projektpartner, stellvertretend für viele Hersteller von Schweißinvertern, stellt dies ein Hemmnis dar, da in diesem Anwendungsgebiet diese modernen Technologien bislang hauptsächlich prototypisch im Einsatz sind. Dieser prototypische Einsatz findet in der Literatur lediglich bei Invertern mit DC-Ausgang Anwendung. Für flexibel einsetzbare Schweißinverter, die einen breiteren Anwendungsbereich abdecken sollen, sind jedoch zusätzlich auch verschiedene AC-Signalformen am Ausgang erforderlich. Daraus ergibt sich auch eine Forschungsfrage, wie die Topologie für diesen Einsatzzweck mit den neuen Schaltelementtechnologien realisiert werden könnte.

Im Gegensatz zu der Verwendung von Wide-Bandgap-Halbleitern ist der Einsatz einer aktiven PFC bei Schweißinvertern bereits häufig anzutreffen. Dies dient der Verbesserung des Leistungsfaktors und der Reduktion von Oberschwingungen in den eingangsseitigen Stromverläufen. Der Einsatz einer aktiven PFC eröffnet außerdem neue Möglichkeiten hinsichtlich der Regelung des Gesamtsystems, welches aus PFC, Hauptinverter und der für Inverter mit AC-Ausgang benötigten Schaltung zur Ausgangssignalformung besteht. Anstatt die einzelnen Baugruppen regelungstechnisch getrennt voneinander zu betrachten, könnte eine gesamtheitliche übergeordnete Regelung der Teilsysteme weiteres Potential für Effizienzsteigerungen bieten.

Der Einsatz von modernen Halbleiterelementen, wie SiC-MOSFETs, bedingt im Vergleich zu aktuell genutzten Technologien erhöhte Bauteilkosten. Diese werden zwar durch eine höhere Effizienz über die Laufzeit im Betrieb relativiert, resultieren jedoch zunächst in einer Steigerung des Gerätepreises. Durch eine ganzheitliche Sichtweise auf alle in einem Schweißinverter typischerweise genutzten Baugruppen können jedoch auch Synergieeffekte genutzt werden, um die Anzahl der notwendigen Schaltelemente zu reduzieren. Eine offene Forschungsfrage dabei ist die Reduktion von Schaltelementen durch die Integration des bislang separat aufgebauten Pilotinverters in die Hauptplatine. Dieser ist in einigen Anwendungen notwendig, um den Lichtbogen aufzubauen bzw. aufrecht zu erhalten.

Ergebnis der im Rahmen des TBI durchgeführten Patentrecherche ist, dass die Entwicklung einer neuen Generation von Schweißinvertern mit den oben genannten Neuerungen

- Einsatz von Wide-Bandgap-Halbleitern,
- Einsatz einer aktiven PFC,
- (teilweise) Integration des Pilotinverters, bzw. dessen Funktion in den Hauptinverter, sowie
- Realisierung einer übergeordneten Regelung des Gesamtsystems anstatt lediglich einer separaten Regelung der einzelnen Baugruppen,

aufgrund bestehender Patente keinen wesentlichen Einschränkungen unterliegt.

Vor allem für KMU ist der mit einer solchen Neuentwicklung einhergehende Forschungs- und Entwicklungsaufwand allein jedoch kaum zu bewältigen. Neben fachlicher Expertise im Leistungselektronikdesign mit Wide-Bandgap-Halbleitern, der Erstellung von Simulationsmodellen und der digitalen Regelung ist auch moderne Laborinfrastruktur erforderlich. Es wurden verschiedene Möglichkeiten zur Beschleunigung des Entwicklungsprozesses erörtert, darunter Hardware-in-the-Loop (HIL)-Tests. Diese werden als eine Möglichkeit betrachtet, um Regelungsalgorithmen und Topologien bzw. Schaltungsdesigns bereits in einem frühen Entwicklungsstadium in einer simulierten, aber realitätsnahen Umgebung zu testen. Dabei wird die Hardware, auf der die Regelungsalgorithmen ausgeführt werden mit einem Echtzeitsystem gekoppelt, welches die zu regelnden Systeme (PFC, Hauptinverter, etc.) emuliert, siehe Abbildung 3.

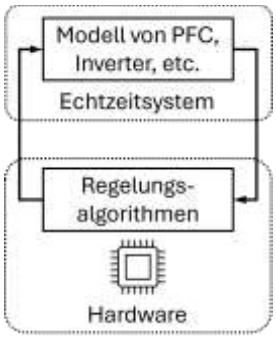

Abbildung 3. Schematische Darstellung einer HIL-Testumgebung.

Diese Methode ermöglicht es, Hardware und Algorithmen frühzeitig und sicher zu validieren, ohne das physische Zielsystem zu benötigen und ohne die Gefahr, dass bei Softwarefehlern Schaden an bereits vorhandenen realen Aufbauten entsteht [11]. Zusätzlich reduziert sich damit der zeit- und ressourcenaufwändige Aufbau unterschiedlicher Prototypen. Das Vorhaben kann zudem von modernen Entwicklungspraktiken wie Clean Code, Design Patterns, Modularisierung sowie der Implementierung einer Continuous-Integration- und Continuous-Deployment-(CI/CD)-Pipeline mit automatisierten Unit-Tests profitieren [12], [13].

Zusammenfassung

Die im Rahmen des TBI durchgeführten Messungen an einem Schweißinverter, wie er oftmals in der Industrie eingesetzt wird, haben gezeigt, dass hinsichtlich Energieeffizienz und Netzrückwirkungen

Verbesserungspotential besteht. Insbesondere der eingangs erwähnte große Anwendungsbereich von Schweißinvertern in der EU zeigt die Bedeutung der Reduktion von Emissionen welche mit Effizienzsteigerungen bei Schweißinvertern einher geht. Um seitens der Hersteller auch auf zukünftige gesetzliche Richtlinien zu Energieeffizienz und Netzrückwirkungen vorbereitet zu sein, wird ein Umstieg auf moderne Wide-Bandgap-Halbleiter, die in anderen Industriefeldern bereits dem Stand der Technik entsprechen, sowie der Einsatz einer aktiven PFC empfohlen. Zudem bietet die Entwicklung eines innovativen Regelungskonzeptes für das Gesamtsystem bestehend aus PFC, Hauptinverter, Funktion des Pilotinverters, und der Schaltung zur Ausgangssignalformung Potential zur Effizienzsteigerung. Der Einsatz moderner digitaler Methoden, wie zum Beispiel Hardware-in-the-Loop-Tests können den Entwicklungsprozess beschleunigen. Im Rahmen des TBI wurde die Grundlage für weitere Forschungsprojekte in Kooperation mit Industriepartnern geschaffen, um die aufgezeigten Verbesserungspotenziale sowie die diskutierten Möglichkeiten zur Ausschöpfung dieser Potenziale praktisch umzusetzen. Aus heutiger Sicht könnte die Realisierung der vorgestellten Lösungen den Weg zu einer Technologieführerschaft auf dem Gebiet der Schweißinverter ebnen, sowie den Gesamtenergiebedarf in diesem Sektor erheblich senken.

Danksagung

Die Autoren bedanken sich bei der Firma SBI GmbH, insbesondere bei J. Niedermayer und F. Spulak, für die Bereitstellung eines dem Industriestandard entsprechenden Schweißinverters samt Peripherie zur Durchführung der Messungen, sowie für den fachlichen Austausch.

Literatur

[1] Verordnung (EU) 2019/1784 der Kommission vom 1. Oktober 2019 zur Festlegung von Anforderungen an die umweltgerechte Gestaltung von Schweißgeräten gemäß der Richtlinie 2009/125/EG des Europäischen Parlaments und des Rates. Zugegriffen: 25. Juli 2025. [Online]. Verfügbar unter: https://eur-lex.europa.eu/eli/reg/2019/1784/oj

[2] B. Shi, A. I. Ramones, Y. Liu, H. Wang, Y. Li, S. Pischinger, und J. Andert, „A review of silicon carbide MOSFETs in electrified vehicles: Application, challenges, and future development“, IET Power Electronics, Bd. 16, Nr. 12, S. 2103–2120, Sep. 2023, doi: 10.1049/pel2.12524.

[3] R. Unseld, „“I estimate that 70 % of the SiC demand will account for the e-mobility sector”“, ATZ Electronics Worldwide, Bd. 19, Nr. 6, S. 14–17, Juni 2024, doi: 10.1007/s38314-024-1886-9.

[4] S. Acharya, N. S. Chauhan, und S. K. Mishra, „Replacing Si-IGBT by SiC Mosfet in high Gain Inverter: Challenges and Opportunities“, in Tagungsband der 2018 IEEE International Conference on Power Electronics, Drives and Energy Systems (PEDES), Chennai, India: IEEE, Dez. 2018, S. 1–6. doi: 10.1109/pedes.2018.8707565.

[5] S. Hain, M. Meiler, und M. Denk, „Evaluation of 800V Traction Inverter with SiC-MOSFET versus Si-IGBT Power Semiconductor Technology“, in Tagungsband der PCIM Europe 2019; International Exhibition and Conference for Power Electronics, Intelligent Motion, Renewable Energy and Energy Management, Nürnberg, Deutschland, Mai 2019, S. 1–6.

[6] M. Nitzsche, C. Cheshire, M. Fischer, J. Ruthardt, und J. Roth-Stielow, „Comprehensive Comparison of a SiC MOSFET and Si IGBT Based Inverter“, in Tagungsband der PCIM Europe 2019; International Exhibition and Conference for Power Electronics, Intelligent Motion, Renewable Energy and Energy Management, Nürnberg, Deutschland, Mai 2019, S. 1–7.

[7] X. Ding, M. Du, T. Zhou, H. Guo, C. Zhang, und F. Chen, „Comprehensive Comparison between Sic-mosfets and Si-igbts Based Electric Vehicle Traction Systems under Low Speed and Light Load“, Energy Procedia, Bd. 88, S. 991–997, Juni 2016, doi: 10.1016/j.egypro.2016.06.124.

[8] A. Marneni, N. Babu, S. K. Subramaniam, P. Murugesan, und J. J. Prakash, „Design, analysis and testing of three-phase isolated SiC-based high current DC-DC converter for arc welding power supply“, Electr Eng, Apr. 2025, doi: 10.1007/s00202-025-03089-5.

[9] J. W. Kolar und T. Friedli, „The Essence of Three-Phase PFC Rectifier Systems—Part I“, IEEE Transactions on Power Electronics, Bd. 28, Nr. 1, S. 176–198, Jan. 2013, doi: 10.1109/tpel.2012.2197867.

[10] T. Friedli, M. Hartmann, und J. W. Kolar, „The Essence of Three-Phase PFC Rectifier Systems—Part II“, IEEE Transactions on Power Electronics, Bd. 29, Nr. 2, S. 543–560, Feb. 2014, doi: 10.1109/tpel.2013.2258472.

[11] H. P. Nguyen, T. T. Nguyen, M. P. Le, und T. N. Tran, „Experimental platforms consisting of software-in-the-loop, hardware-in-the-loop and power-hardware-in-the-loop for developing MPPT charge controller for photovoltaic system“, Solar Energy, Bd. 298, S. 113666, Sep. 2025, doi: 10.1016/j.solener.2025.113666.
[12] A. Barbie, W. Hasselbring, und N. Pech, „Continuous Integration Testing of Embedded Software with Digital Twin Prototypes“, 26. Juni 2021, TechRxiv. doi: 10.36227/techrxiv.14770983.v1.
[13] P. Cassieri, M. Esposito, S. Romano, und G. Scanniello, „On the use of Test-Driven Development for Embedded Systems“, Information and Software Technology, Bd. 187, S. 107779, Nov. 2025, doi: 10.1016/j.infsof.2025.107779.

Multi-Sensor-Imaging für die qualitativen Charakterisierung von gebrauchten Solarmodulen

Alexander Eckert
Austrian Institute of Technology
Alexander.eckert@ait.ac.at

Georg Kaniak
Austrian Institute of Technology
Georg.kaniak@ait.ac.at

Philipp Jelke
Austrian Institute of Technology
Philipp.jelke@ait.ac.at

Laurin Ginner
Austrian Institute of Technology
Laurin.ginner@ait.ac.at

Kurzzusammenfassung

Die kosteneffiziente Qualitätsprüfung gebrauchter Solarmodule für eine mögliche Wiederverwendung wird angesichts stark steigender Rücklaufmengen zunehmend wichtiger. Ein niederösterreichisches Startup beauftragte das Austrian Institute of Technology (AIT) daher mit der Entwicklung eines vollautomatisierbaren Multi-Sensor-Imaging-Inspektionssystems. Ziel war die Konzeption und Erprobung einer optischen Inspektionsstation. Dafür wurden verschiedene Beleuchtungssysteme, Kameras und Objektive unter realen Bedingungen getestet und verglichen. Auf dieser Basis wurden aufgabenoptimierte Systeme implementiert, die nun die Grundlage für eine automatisierte Qualitätsbewertung gebrauchter Solarmodule bilden.

Keywords: Kreislaufwirtschaft, Optische Inspektion

Abstract

The cost-efficient quality assessment of used PV modules for potential reuse is becoming increasingly important due to the expected surge in return volumes. To address this, a Lower Austrian startup commissioned AIT to develop a fully automatable multi-sensor imaging inspection system. The project aimed to support the design and validation of an optical inspection station. For this purpose, various lighting systems, cameras, and lenses were tested and compared under real-world conditions. Based on these evaluations, task-optimized systems were implemented, forming the foundation for automated quality assessment of used PV modules.

Keywords: Circular economy, optical inspection

Einleitung

Weltweit sind inzwischen mehr als 4 Milliarden Solarmodule (Photovoltaikmodule) installiert – allein im letzten Jahr kam eine Milliarde hinzu. Entsprechend wird in den kommenden Jahren ein stark wachsender Rücklauf stillgelegter Photovoltaikmodule (PV-Module) erwartet. Viele davon sind nicht defekt, sondern werden im Zuge von Repowering-Maßnahmen ersetzt, obwohl sie häufig noch Wirkungsgrade von 80–90 % erreichen und lediglich jährliche Leistungsverluste von 1–2 % zeigen [1]. In Deutschland allein wird bis 2030 mit rund 50 Millionen gebrauchten Modulen gerechnet, die aufbereitet werden müssen [2]. Diese Entwicklung verdeutlicht den dringenden Bedarf an effizienten Wiederverwendungs- und Recyclingstrategien – und eröffnet zugleich Chancen für automatisierte Lösungen im Bereich Kreislaufwirtschaft.

Die derzeitigen Qualitätsprüfungen gebrauchter PV-Module erfolgen überwiegend manuell. Da visuelle und elektrische Inspektionen von Menschen durchgeführt werden, sind sie zeitaufwändig, uneinheitlich und schwer skalierbar – und stellen damit einen zentralen Engpass im Wiederverwendungsprozess dar. Trotz des großen Marktbedarfs ist der Automatisierungsgrad in der Modulinspektion bislang gering. Momentan werden zwar nahezu alle Rückläufer visuell begutachtet, ein hoher Anteil wird darüber

hinausgehend aber keiner weiteren Inspektionen unterzogen [3]. Der überwiegende Teil wird zurzeit ohne genaue Einzelstückcharakterisierung dem stofflichen Recycling zugeführt [4].

Um diese Lücke zu schließen, initiierte das niederösterreichische Startup gemeinsam mit dem AIT ein Entwicklungsprojekt. Ziel war es, manuelle Prüfungen durch kamerabasierte, automatisierte Inspektionssysteme zu ersetzen. Der Fokus lag auf der Entwicklung einer robusten, hochdurchsatzfähigen optischen Inspektionslösung, die Defekte zuverlässig erkennt, den manuellen Aufwand reduziert und so skalierbare Wiederverwendungswege ermöglicht.

Problemlösungsweg / Vorgehen

Zur Entwicklung eines robusten, anwendungsspezifischen Inspektionssystems analysierte das AIT zunächst typische gebrauchte PV-Module, gängige Defektmuster sowie deren Erkennungsmöglichkeiten [5]. Relevante Schadensbilder umfassten u. a. Risse, inaktive Zellen, Delamination und Blasenbildung auf der Rückseite.

Unter Berücksichtigung mechanischer Randbedingungen, hoher Verarbeitungsgeschwindigkeit und notwendiger optischer Auflösung wurde ein maßgeschneidertes Prüfsystem konzipiert. Dieses kombiniert mehrere Kamera- und Beleuchtungskonfigurationen:

- Vorderseiteninspektion: Hochauflösende Elektrolumineszenz-Kameras (EL) [6] zur Erkennung elektrischer und struktureller Defekte, ergänzt durch RGB-Kameras zur visuellen Oberflächenbewertung.
- Rückseiteninspektion: RGB-Kameras in Kombination mit Linienleuchten und gerichteter Beleuchtung zur Hervorhebung von Oberflächentexturen und zur Detektion von Delamination oder Blasenbildung.

Die resultierende Multi-Sensor-Konfiguration erlaubt eine aufgabenoptimierte, automatisierte Prüfung mit hohem Durchsatz und wurde gezielt für die Anforderungen der geplanten Inspektionslinie entwickelt.

Ergebnis

Im Projekt wurde ein Multikamera-Inspektionssystem erfolgreich entwickelt, getestet und auf die Anforderungen einer automatisierten PV-Modul-Inspektionslinie zugeschnitten. Die Evaluation zeigte, dass unterschiedliche Inspektionskriterien den Einsatz verschiedener Kameras, Objektive und Beleuchtungssysteme erfordern. Die ausgewählten Komponenten wurden durch Tests validiert und anschließend für die Integration empfohlen, siehe Abbildung 1.

Das System ist mittlerweile in der Produktionsumgebung des Auftraggebers im Einsatz und ermöglicht eine zuverlässige Hochgeschwindigkeitsinspektion. Damit wurde ein wichtiger Schritt hin zur vollautomatischen Qualitätsprüfung gebrauchter PV-Module erreicht.

Dieses Ergebnis stärkt nicht nur die technologische Basis der Kreislaufwirtschaft im Photovoltaiksektor, sondern kann auch auf andere Akteure in Europa und darüber hinaus übertragen werden. Darüber hinaus bietet der Multi-Sensor-Ansatz Potenzial zur Erweiterung auf andere komplexe Materialströme im Bereich des Recyclings.

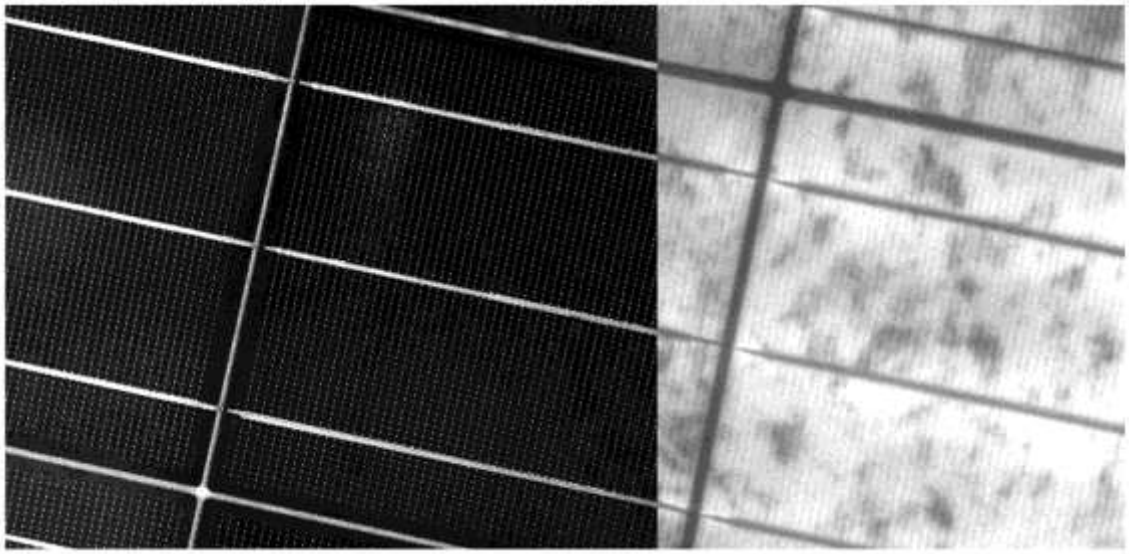

Abbildung 1: Polykristalline Solarzelle die mit verschiedenen Kamerasystemen aufgenommen wurde (links: RGB, rechts: monochrom, Elektrolumineszenz)

Zusammenfassung

Da sich Milliarden von PV-Modulen ihrem Lebensende nähern, wird eine effiziente Wiederverwendung immer wichtiger. Viele ausgemusterte Module sind weiterhin funktionsfähig, doch die derzeitigen manuellen Prüfverfahren sind zu langsam und arbeitsintensiv. Gemeinsam mit dem AIT entwickelte das niederösterreichische Startup daher ein automatisiertes Multikamera-Inspektionssystem. Durch die Kombination von EL- und RGB-Kameras mit spezialisierter Beleuchtung ermöglicht es eine schnelle und zuverlässige Qualitätsbewertung. Das System ist bereits in die Inspektionslinie integriert und leistet einen entscheidenden Beitrag zur Skalierbarkeit der Wiederverwendung und zur Umsetzung von Kreislaufwirtschaftsmodellen im Photovoltaiksektor.

Literatur

[1] Santos, Sofia & Torres, João Paulo & Fernandes, Carlos & Lameirinhas, Ricardo. (2021). The impact of aging of solar cells on the performance of photovoltaic panels. Energy Conversion and Management: X. 10. 100082. 10.1016/j.ecmx.2021.100082.

[2] IRENA and IEA-PVPS (2016), “End-of-Life Management: Solar Photovoltaic Panels,” International Renewable Energy Agency and International Energy Agency Photovoltaic Power Systems, ISBN 978-92-95111-99-8

[3] Umweltbundesamt (2024, 26. März). Photovoltaik – Behandlung und Recycling. Abgerufen von Umweltbundesamt-Webseite

[4] IEA PVPS Task 12 (2022). Status of PV Module Recycling in Selected IEA PVPS Task 12 Countries. International Energy Agency Photovoltaic Power Systems Programme

[5] Maziuk, Maurycy & Jasińska, Laura & Domaradzki, Jaroslaw & Chodasewicz, Paweł. (2023). Imaging methods of detecting defects in photovoltaic solar cells and modules: a survey. Metrology and Measurement Systems. 30. 10.24425/mms.2023.146426.

[6] Frazão, M. & Silva, José & Lobato, Killian & Serra, J.M.. (2016). Electroluminescence of silicon solar cells using a consumer grade digital camera. Measurement. 99. 10.1016/j.measurement.2016.12.017.

Digitalisierung eines Energieeffizienzkonzepts mit Reportingfunktion

Clemens Schwaiger
CDP Center for Digital Production GmbH
clemens.schwaiger@acdp.at

Sascha Gent
CDP Center for Digital Production GmbH
sascha.gent@acdp.at

Kurzzusammenfassung

Das Projekt konzentrierte sich auf die Überwachung des Energieverbrauchs in einem Walzwerk. Es ermöglicht eine Analyse des Energieaufwands pro Produkt und je Gewicht an eingesetztem Rohmaterial. Zudem werden die Energieverbräuche in drei Betriebszuständen der Produktionsanlagen eingeteilt und auf einem Dashboard visualisiert. Dieses dient sowohl der Verbesserung der Energieeffizienz als auch der Fehlererkennung, indem es Anomalien im Verbrauch aufzeigt und frühzeitig auf technische Störungen hinweist. So können Produktionsprozesse optimiert und Ausfallzeiten minimiert werden, wodurch eine nachhaltige und kosteneffiziente Fertigung unterstützt wird. Die Ergebnisse werden den Anwendern aufbereitet zur Verfügung gestellt und in ein bestehendes Supervisory Control and Data Acquisition (SCADA) - System integriert.

Keyword: Digitalisierung, Energieeffizienz, Kalander

Abstract

The project focused on energy monitoring in a rolling mill by enabling the analysis of energy consumption per product and per weight of raw material used. Furthermore, this energy consumption is divided into three operating states of the production facilities and visualized on a dashboard. This serves both to improve energy efficiency and detect errors by highlighting consumption anomalies and providing early warning of technical malfunctions. It allows production processes to be optimized and downtime minimized, thus supporting sustainable and cost-efficient production. The results are processed and made available to users and integrated into an existing SCADA system.

Keyword: Digitalization, energy efficiency, Kalander

Einleitung

Gegenstand dieses Beitrags ist ein Walzwerk welches aus mehreren Anlagen zur Glättung und Verdichtung von Gummi besteht. Diese „*Kalander*“ stellen eine zentrale Komponente in der Produktion von Dichtungsmaterial dar. Sie bestehen grundsätzlich aus zwei polierten und beheizten Walzen, die eine variable Spalte bilden. Durch diese wird das Material in mehreren Durchgängen geführt, um eine Matte mit der gewünschten Stärke zu fertigen.

Jene Kalander werden mit elektrischer Energie zum Betrieb der Antriebe und mit Wasserdampf zur Temperierung der Walzen versorgt. Die Peripherie umfasst eine Absauganlage und eine Lösemittelrückgewinnungsanlage, wobei zusätzlicher Bedarf an elektrischer Energie und Wasserdampf entsteht. Ein Überblick der Energieflüsse in und um das Walzwerk ist in Abbildung 1 dargestellt.

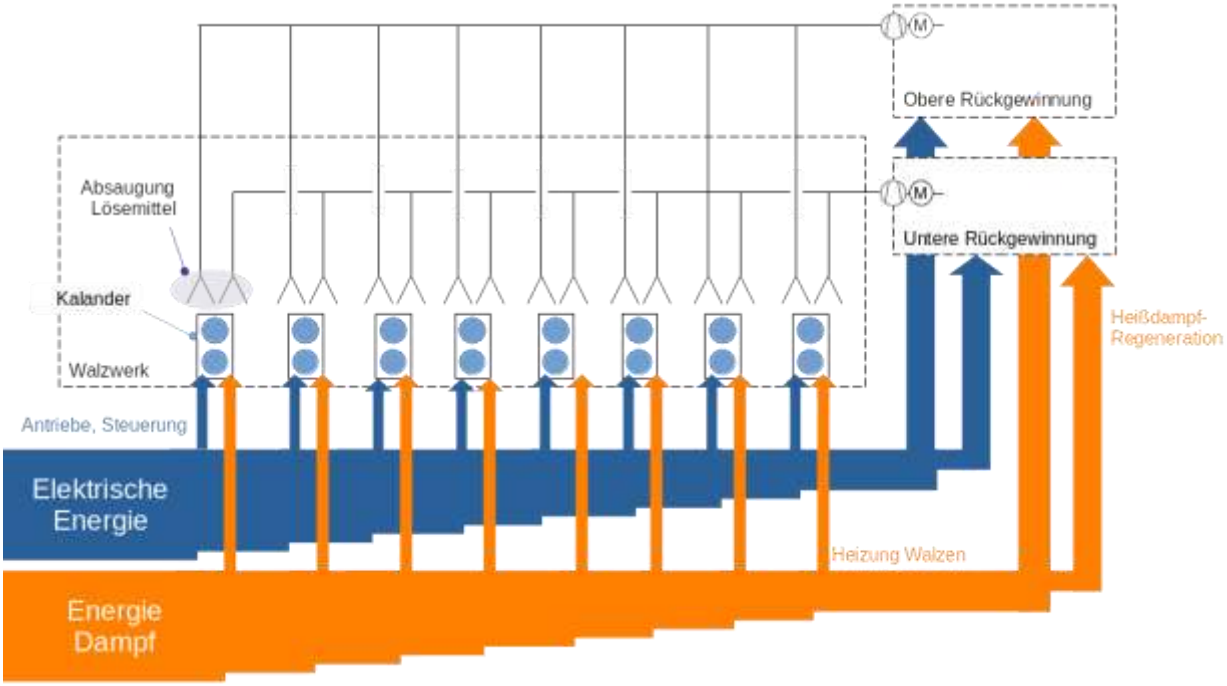

Abbildung 1. Energieflüsse im Bereich Walzwerk

Insbesondere durch den Verbrauch von Dampf, welcher durch die Verbrennung fossiler Energieträger hergestellt wird, stellen diese Kalander energieintensive Produktionsmaschinen dar. Die Optimierung des Energieverbrauchs ist somit ein wesentlicher Kostenfaktor und wirkt sich darüber hinaus positiv auf die CO2-Bilanz aus.

Energiemonitoring und -management, z.B. in Form gesetzlich verpflichtender Managementsysteme wie ISO 50001, stellen die Basis für Verbesserungsmaßnahmen dar [1]. Derartige Strukturen müssen jedoch erweitert werden, um tatsächlich strategisches Energiemanagement zu ermöglichen [2]. Konkret bedeutet dies die Harmonisierung und Kontextualisierung von Energiedaten mit anderen Produktionsdaten zur Schaffung einer informativen Datenbasis. Nur so kann der nötige Informationsgehalt für Entscheidungsträger und aufbauende Analysen bereitgestellt werden. Eine solche Datenbasis war vor diesem Projekt im Unternehmen noch nicht vorhanden, weswegen die Energieverbräuche abgesehen von absoluten Werten weitestgehend undurchsichtig waren und somit Optimierungen nur schwer umsetzbar waren. Am Markt erhältliche Produkte (z.B. FourJaw, iFactory und ähnliche) können den sehr spezifischen Produktionsablauf nicht ausreichend abbilden oder sind – wie der Marktführer EnMS von ABB – zu komplex um sie in ein KMU integrieren zu können. Dies würde zu einem Overhead im Personal sorgen, der mit den angestrebten Einsparungen nicht aufzuwägen wäre. Aus diesem Grund war eine maßgeschneiderte Lösung, die zusätzlich mit den bereits vorhandenen Systemen im Unternehmen kommunizieren kann die einzige Lösung. Jene Lösung zur Ermittlung und Aufbereitung der benötigten Daten für ein aussagekräftiges und robustes Energiemonitoring ist Inhalt des Projekts, das diesem Beitrag zu Grunde liegt.

Problemlösungsweg / Vorgehen

Die Kontextualisierung erfolgt spezifisch für die jeweilige Produktion und muss anhand der Arten der Energieformen, Anlagen, Materialien und Betriebsstrategien erfolgen. Hierbei können unterschiedliche Ansätze verfolgt werden. Im Fall des dieses Walzwerkes sollen folgende Kennzahlen abgeleitet werden:

- Energieverbrauch bezogen auf das Gewicht des eingesetzten Materials
- Energieverbrauch bezogen auf einzelne Produkte
- Energieverbrauch je Betriebszustand der Anlage
- Aufschlüsselung der Energieformen

Der Energieverbrauch je Betriebszustand erfordert die Definition des Betriebszustandes. Ein abstrakter und universeller Zugang ist die Einteilung der Zustände gemäß ihrer Wertschöpfung:

a. Direkt wertschöpfend: Unmittelbar für den Produktionsprozess nötiger Energieeinsatz (z.B.: Energie zum Antrieb der Walzen während des Walzvorganges)
b. Indirekt wertschöpfend: Sonstiger Energieverbrauch der für die Produktion erforderlichen Vorgänge. (z.B.: Energie zum Betrieb der Gebläse zum Absaugen von Lösemitteldampf)
c. Nicht wertschöpfend: (Theoretisch) vermeidbare Zustände
(z.B.: Betrieb des Hydraulikaggregats zur Aufrechterhaltung des Systemdrucks in (längeren) Zeiten ohne Herstellungszyklen)

Die benötigten Energie- und Produktionsdaten wurden im hauseigenen SCADA System zusammengeführt und die nötigen Berechnungen mittels Python und SQL durchgeführt. Die kumulierte Energie pro Herstellungszyklus je Produkt/Gewicht liefert diskrete Werte. D.h. je Kennzahl und Herstellungszyklus wird ein Wert in der Datenbank abgelegt. Die Einteilung gemäß Wertschöpfung hingegen liefert Zeitreihen, die gemäß obiger Definition in Segmente eingeteilt werden. Diese Zeitreihen können ihrerseits für gewisse Zeitfenster ausgewertet werden (z.B. wertschöpfende Vorgänge je Monat).

Ergebnis

Im Rahmen des Projekts wurde eine umfassende Lösung zur Digitalisierung des Energieeffizienzkonzepts in einem energieintensiven Walzwerk entwickelt und implementiert. Die erfassten und aufbereiteten Informationen werden über ein speziell entwickeltes Dashboard im firmeneigenen SCADA-System bereitgestellt (siehe Abbildung 2). Zu Beginn wählt der Anwender das relevante Zeitfenster sowie einen bestimmten Kalander aus. Daraufhin erfolgt eine detaillierte Darstellung aller Chargen, die in diesem Zeitraum bearbeitet wurden. Nach dem Start des Analyseprozesses berechnet das System verschiedene Kennzahlen, die eine differenzierte Analyse der Energieverbräuche ermöglichen. Die Visualisierungen umfassen unter anderem den Energieverbrauch pro Charge und pro eingebrachtem Materialgewicht. Besonders hervorgehoben wird dabei die Unterscheidung zwischen wertschöpfenden und nicht wertschöpfenden Energieverbräuchen, was eine präzise Identifikation von Optimierungspotenzialen innerhalb der Produktionsprozesse erlaubt. Ergänzend werden die Energieverbräuche einzelner Walzvorgänge in Form eines Boxplots dargestellt, wodurch sich Unregelmäßigkeiten und Ausreißer im Energiebedarf einzelner Produktionsschritte identifizieren lassen. Die zugrundeliegenden Zeitreihen zu den unterschiedlichen Wertschöpfungsbeiträgen werden für jeweils gewählte Zeitfenster analysiert und bieten so einen strukturierten Überblick über den Verlauf und die Effizienz der Produktion.

Das Dashboard unterstützt durch seine klare und adaptierbare Darstellung die gezielte Überwachung sowie den Vergleich von Energieverbrauchsdaten verschiedener Chargen und Produktvarianten. Auffällige Abweichungen werden unmittelbar sichtbar, sodass sowohl technische Störungen als auch ungewöhnliche Energieverbräuche erkannt werden können. Diese Erkenntnisse ermöglichen eine zielgerichtete Fehlerdiagnose und bieten darüber hinaus Ansatzpunkte für weitere Maßnahmen zur Steigerung der Energieeffizienz.

Abbildung 2. Dashboard

Ein zentraler Fokus des Projekts war die universelle Anwendbarkeit der entwickelten Systemarchitektur und Methodik. Sowohl das System selbst als auch die gewählten Ansätze zur Erhebung und Kontextualisierung der Daten wurden modular konzipiert, sodass sie mit vergleichsweise geringem Aufwand auf andere Produktionslinien oder Industriebereiche übertragen werden können. Die Definition der Kennzahlen wie auch die Segmentierung der Betriebszustände lassen sich flexibel an unterschiedliche technische Rahmenbedingungen und Branchenanforderungen anpassen. Darüber hinaus ermöglichen standardisierte Schnittstellen eine problemlose Integration in andere SCADA- oder Betriebsführungssysteme und gewährleisten so eine breite Kompatibilität der erarbeiteten Lösung. Auch die Visualisierungsmechanismen und Auswertungstools können ohne grundlegende Änderungen an neue Produktionsprozesse, Maschinenarten oder Berichtsstrukturen angepasst werden. Insbesondere die Methoden zur Identifikation und zum Reporting von Anomalien im Energieverbrauch erweisen sich als universell einsetzbar und sind für verschiedenste industrielle Anwendungen direkt anschlussfähig. Diese ganzheitliche Betrachtung und Vorgehensweise fördert nicht nur die Optimierung und Fehlerreduktion im betrachteten Walzwerk, sondern schafft zudem die Voraussetzungen für eine nachhaltige Generalisierung und Übertragbarkeit auf andere industrielle Kontexte, unabhängig von Branche, Produkt oder Unternehmensgröße.

Zusammenfassung

Der Beitrag beschäftigt sich mit einem energieintensiven Walzwerk, das in der Produktion von Dichtungsmaterialien eingesetzt wird. Zentrale Komponenten sind Kalander bestehend aus zwei beheizten Walzen, die Rohmaterial glätten und verdichten. Der Betrieb erfordert elektrische Energie sowie Dampf, letzterer meist aus fossilen Quellen, was die Relevanz einer Energieoptimierung erhöht. Zur Energieeinsparung und Verbesserung der CO_2-Bilanz wird das bestehende Energiemonitoring und -management erweitert, um die Kontextualisierung von Energiedaten mit Produktionsdaten zu ermöglichen.

Dabei werden spezifische Kennzahlen ermittelt, z.B.:

- Energieverbrauch pro Materialgewicht
- Energieverbrauch je Produkt
- Energieverbrauch je Betriebszustand (direkt, indirekt, nicht wertschöpfend)
- Aufschlüsselung nach Energieformen

Die konzeptionelle Modularität und die verwendeten Kennzahlen sind zudem mit minimalem Aufwand auf andere Produktionsprozesse und Maschinen übertragbar. Dadurch können sowohl große Industrieunternehmen als auch kleinere Produktionsbetriebe von den Ansätzen und dem System profitieren. Schnittstellen und Visualisierungsmöglichkeiten sind universell gestaltet und lassen sich flexibel an unterschiedliche IT- und Produktionsumgebungen anpassen.

Die Analyse erfolgt durch Datenverarbeitung mit Python und SQL im bestehenden SCADA-System. Die Ergebnisse werden in einem Dashboard visualisiert, das Energieverbrauchsdaten je Charge und Prozessschritt aufzeigt – auch unterteilt nach deren Wertschöpfungsbeitrag. Das System hilft sowohl bei der Energieoptimierung als auch bei der Fehlererkennung in der Produktion, etwa durch Identifikation von Ausreißern beim Energieverbrauch ähnlicher Chargen.

Literatur

[1] International Organization for Standardization, 23 07 2025. [Online]. Available: https://www.iso.org/standard/69426.html.

[2] M. Hacksteiner, G. Fuchs und F. Bleicher, „Strategic energy management in mechanical series production: An industrial use-case,“ *Procedia Manufacturing,* 2019.

Gießprozess für Sandformen mit dynamischer Vakuumsteuerung ROBOCAST V

Edmundo Oliveira
Fill G.m.b.H
Edmundo.oliveira@fill.co.at

Sophie Binder
Fill G.m.b.H
Sophie.binder@fill.co.at

Kurzzusammenfassung

In diesem Beitrag wird eine innovative Simulationsstrategie für das ROBOCAST V-System im Rahmen einer TBI-Initiative vorgestellt. Mithilfe der Kombination von Finite Element Method (FEM) und Computational Fluid Dynamics (CFD) sowie zeitgesteuerten Widerstandsfiltern konnten dynamische Gießprozesse effizient und ressourcenschonend modelliert werden. Dies ermöglichte eine risikominimierte, datenbasierte Bewertung von Prozessflexibilität, Gussqualität und Systemeffizienz und zeigt den Nutzen digitaler Simulationen für die Optimierung industrieller Gießverfahren auf.

Keywords: Füllsimulation, dynamisches Netz, Filter-Trick

Abstract

This paper presents an innovative simulation strategy for the ROBOCAST V system as part of a TBI initiative. Using a combination of FEM (finite element method) and CFD (computational fluid dynamics) as well as time-controlled resistance filters, dynamic casting processes could be modeled efficiently and, in a resource-saving manner. This enabled a risk-minimized, data-based evaluation of process flexibility, casting quality, and system efficiency, demonstrating the benefits of digital simulations for optimizing industrial casting processes.

Keywords: Filling simulation, dynamic network, filtering trick

Einleitung

Im sich rasant entwickelnden Umfeld der modernen Gießereitechnologie stehen Prozesseffizienz und Gussqualität zunehmend im Mittelpunkt. Eine deutsche Gießerei hat das transformative Potenzial digitaler Simulation erkannt und ein TBI-Projekt initiiert, um das ROBOCAST V Gießsystem vor einer Anschaffung umfassend zu evaluieren. Dieser Ansatz ermöglicht es dem Unternehmen, die Vorteile und Integrationsmöglichkeiten des Systems in einer risikominimierten, virtuellen Umgebung zu prüfen – statt einer sofortigen physischen Implementierung.
Durch den Einsatz fortschrittlicher Simulationswerkzeuge, darunter die Finite-Elemente-Methode (FEM) und Computational Fluid Dynamics (CFD), wurden kritische Parameter des dynamischen Gießprozesses von ROBOCAST V im Kontext spezifischer Produktionsanforderungen untersucht. Mithilfe digitaler Modelle sollten Erkenntnisse über Prozessflexibilität, Gussqualität und Betriebseffizienz gewonnen werden, um datenbasierte Investitionsentscheidungen zu unterstützen. In diesem Beitrag werden die technischen Methoden sowie die Herausforderungen dargestellt, die während der Simulation aufgetreten sind. Außerdem wird der Mehrwert innovativer digitaler Strategien für die Abbildung und Optimierung komplexer Gießprozesse beleuchtet. In den folgenden Abschnitten werden die eingesetzten Simulationstechniken, die besonderen Fähigkeiten des ROBOCAST V dargestellt in Abbildung 1 und die praktischen Auswirkungen auf industrielle Gießereibetriebe näher erläutert.

Abbildung 1. ROBOCAST V

Problemlösungsweg

In Zusammenarbeit mit MAGMASOFT, einem führenden Anbieter von Gießsimulationssoftware, wurden alternative Strategien für die Simulation im Stand der Technik untersucht. MAGMASOFT verwendet in der Sandgusssimulation eine Methode, bei der eine dünne feste Platte aus demselben Material am Boden des Eingusstrichters platziert wird. Diese Platte schmilzt allmählich und ermöglicht so einen kontrollierten Fluss in die Anschnitte und Hohlräume. Um dieses Prinzip nachzubilden, implementierte MAGMASOFT ein Filtermodell, das als hochwiderstandsfähige Barriere (ähnlich einer Wand) fungiert und zeitgesteuert umgeschaltet werden kann. Im „ON"-Zustand verhindert der Filter den Materialfluss; im ausgeschalteten Zustand („OFF") kann das flüssige Metall ungehindert fließen.

Dieses bekannte Prinzip wurde genutzt, um die Bewegung des ROBOCAST V-Gießrohres zu simulieren. Die kontinuierliche Geometrie des Füllrohres wurde in mehrere kleine Segmente unterteilt (siehe Abbildung 2), wobei jedem Segment eine eigene zeitbasierte Steuerungsfunktion zugewiesen wurde (siehe Abbildung 3).

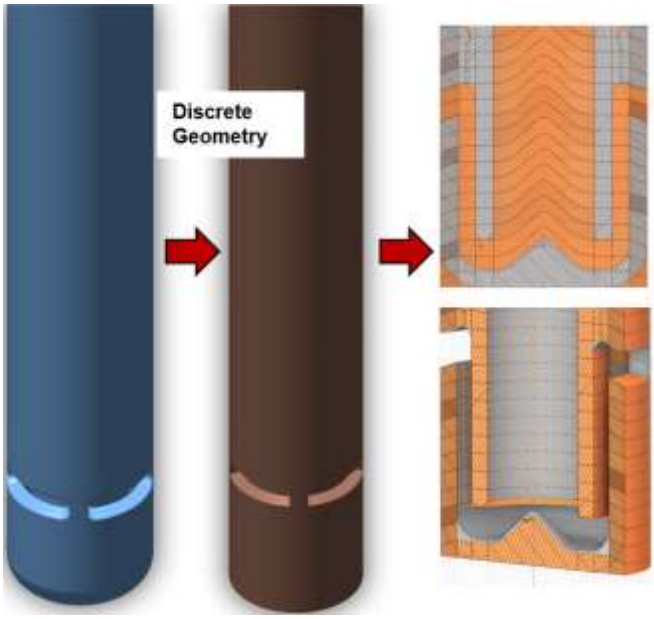

Abbildung 2. Diskrete Geometrie

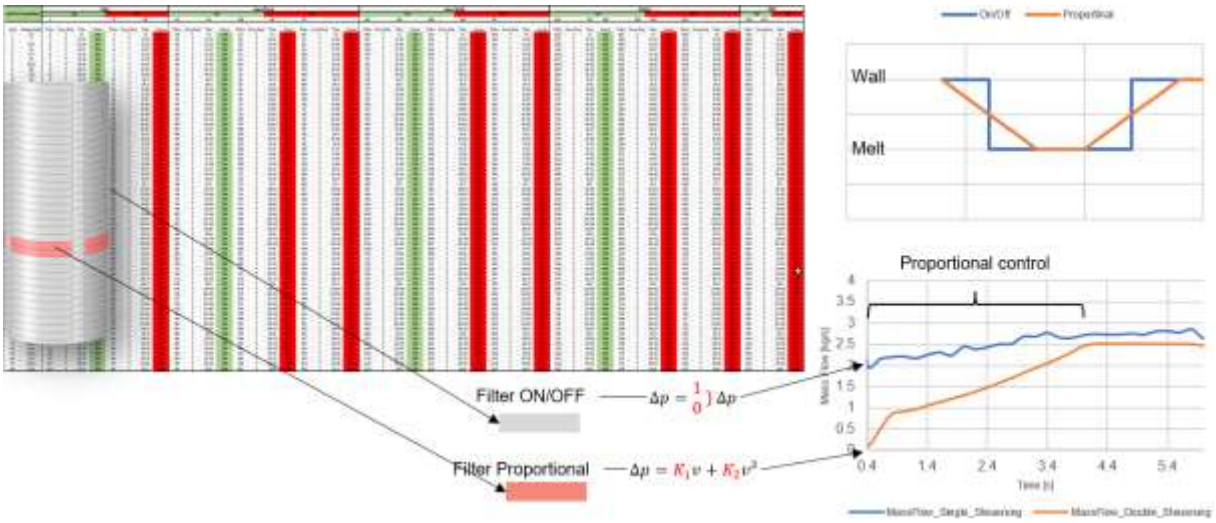

Abbildung 3. Steuertabelle (Filter-Trick)

Durch diese Methode wurde die Notwendigkeit einer dynamischen Netzverfeinerung (Remeshing) eliminiert, während gleichzeitig akzeptable Simulationszeiten für komplexe Gussgeometrien beibehalten wurden. MAGMASOFT erlaubt keine vorgefüllten Volumina als Anfangsbedingung, was bedeutet, dass das Füllrohr während der Füllsimulation erst vollständig gefüllt werden muss, bevor die Formkavität gefüllt werden kann. Dies stellte eine zusätzliche Herausforderung dar: Die Filter am Ausgang des Rohres (in Abbildung 3 pink dargestellt) mussten geschlossen bleiben, bis das Rohr komplett voll war. Ein abruptes Öffnen führte zu unrealistischen Druckspitzen, wie in Abbildung 4 gezeigt.

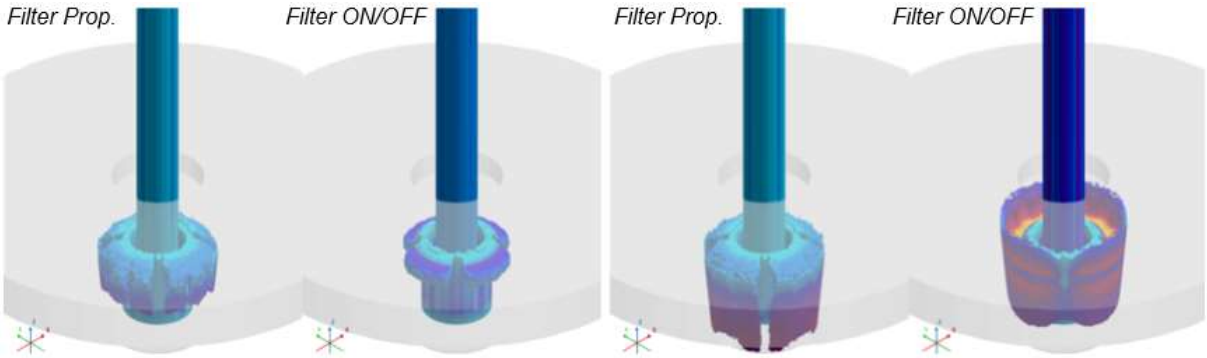

Abbildung 4. Überschwingen durch Druckspitze

Um dieses Problem zu lösen, wurde eine progressive Filtersteuerung implementiert. Durch Reverse Engineering wurden Konstanten (K1, K2) für die Druckverlustgleichung der Filter bestimmt und eine stufenweise Anpassung („ramped adjustment“) anstelle eines plötzlichen Wechsels (0, 1) angewendet, um das allmähliche Ansteigen des Massenstroms, wie in realen Prozessen beobachtet, nachzubilden (siehe Abbildung 4). Der Einsatz der progressiven Filtersteuerung verbessert zudem die Modellierbarkeit der Metallinjektion.

Ergebnis

Als Ergebnis konnten Simulationszeiten erzielt werden, die für industrielle Anwendungen geeignet sind, ohne dabei die Genauigkeit zu beeinträchtigen. Dieser Ansatz zeigte eine hohe Übereinstimmung mit realen Prozessen (siehe Abbildung 5).

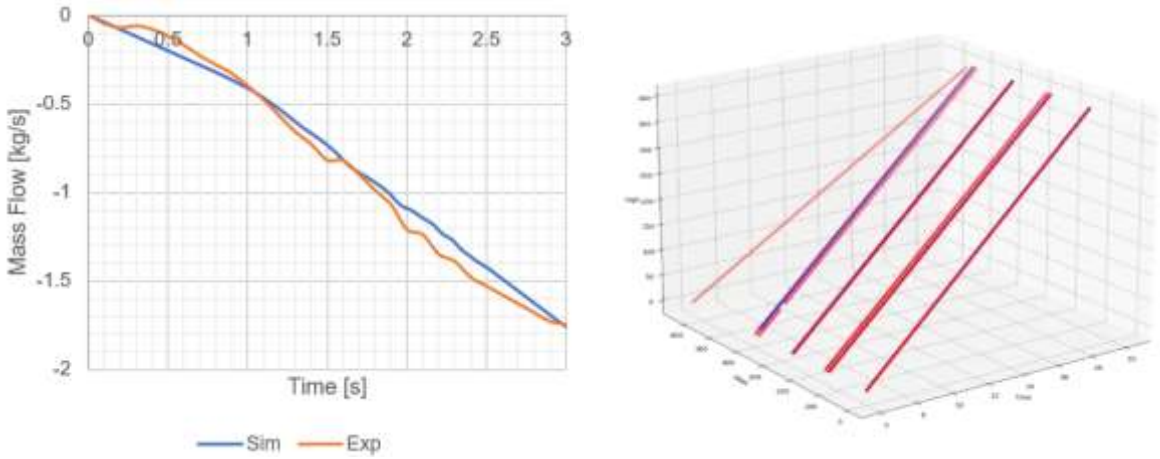

Abbildung 5. Vergleich des Massenstroms und Filterstatus

Die Konfiguration der Filtersteuerung dauert nur wenige Minuten und kann bereits vor der eigentlichen Simulation validiert werden. Das Steuerdiagramm (Abbildung 5) zeigt den ON/OFF-Status, die Aktivierungszeitpunkte, die Filter-ID sowie die jeweilige Z-Achsen-Position an. Abbildung 6 zeigt die Füllsimulation unter Verwendung des dynamischen Füllrohres des ROBOCAST V.

Das Modell hat so gute Ergebnisse geliefert, dass MAGMASOFT ein spezielles Modul entwickelt und in die neue Version 6.1 integriert hat. Das Modul wurde vereinfacht und die Benutzeroberfläche der Software dadurch intuitiver gestaltet. Somit kann jeder MAGMASOFT-Anwender, der seinen Prozess mit ROBOCAST V simulieren möchte, dieses Modul als zusätzliche Funktion erwerben.

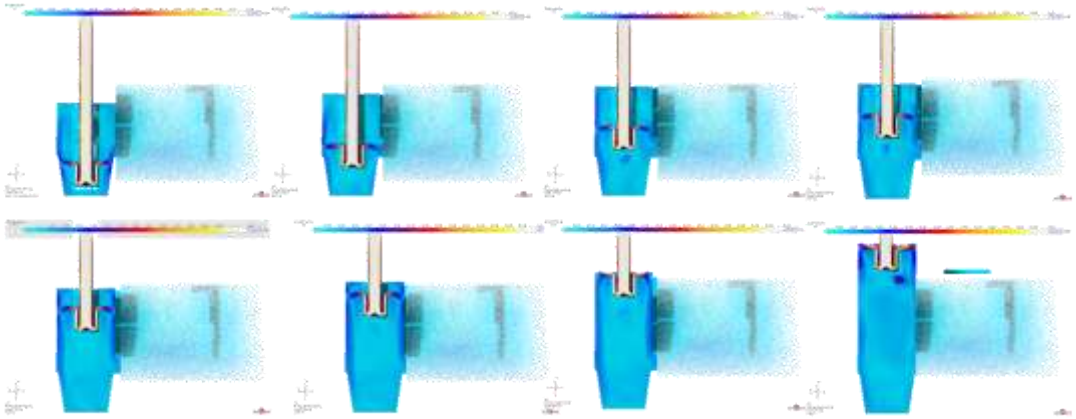

Abbildung 6. Füllsimulation

Zusammenfassung

Dieser Beitrag stellt eine praxisorientierte Methode zur Simulation komplexer Aluminiumguss-Geometrien mit dem ROBOCAST V-System und der MAGMASOFT-Software vor. Die entwickelte progressive Filtersteuerung ersetzt rechenintensive Remeshing-Ansätze und nutzt mittels Reverse Engineering ermittelte Druckverlustkonstanten, um das stufenweise Ansteigen des Massenstroms realitätsnah abzubilden. Herausforderungen wie die Befüllung des Füllrohres vor der Formkavität und die Vermeidung plötzlicher Druckspitzen wurden adressiert. Visualisierungstools und Kontrollplots ermöglichen eine schnelle Konfiguration und Validierung, während die Simulation eine hohe Übereinstimmung mit praktischen Prozessen zeigt. Die Integration eines dedizierten ROBOCAST V-Moduls in MAGMASOFT Version 6.1 vereinfacht die Anwendung und steigert die Zugänglichkeit, sodass industrielle AnwenderInnen effiziente und präzise Prozesssimulationen durchführen können, was vorher nicht möglich war. Alle potentiellen AnwenderInnen (Gießereibetriebe, OEMs, Forschungsinstitute etc.) eines ROBOCAST V oder ähnlichen Gießsystems profitieren von dieser Entwicklung.

CFD-Simulation als Digitales Service für Entwicklung und Produktion

Judith Heim
Linz Center of Mechatronics GmbH
judith.heim@lcm.at

Markus Schörgenhumer
Linz Center of Mechatronics GmbH
markus.schoergenhumer@lcm.at

Martin Trinkl
Linz Center of Mechatronics GmbH
martin.trinkl@lcm.at

Johannes Klinglmayr
Linz Center of Mechatronics GmbH
johannes.klinglmayr@lcm.at

Alexander Dorninger
Linz Center of Mechatronics GmbH
alexander.dorninger@lcm.at

Daniel Polaschek
HAUSER GmbH
Daniel.Polaschek@hauser.com

Kurzzusammenfassung

Numerische Simulationen gewinnen in der industriellen Produktentwicklung zunehmend an Bedeutung. Ihre Anwendung erfordert bislang jedoch umfangreiche Fachkenntnisse und spezialisierte Software- und Hardware-Infrastruktur. Das stellt speziell für kleine und mittlere Unternehmen (KMU) eine Hürde dar. In diesem Beitrag wird ein cloudbasierter Webservice am Beispiel einer Strömungssimulation (CFD-Simulation) für Kühlmöbel vorgestellt. Er wurde entwickelt, um die Bereitstellung von Simulationen in einer Cloudumgebung via Web-Interface zu ermöglichen. Angewendet und getestet wurde der Webservice im Entwicklungs- und Optimierungsprozess von Kühlmöbeln durch Ingenieur:innen, die keine ausgewiesenen CFD-Simulationsexperten sind. Um dies ohne die Anschaffung von teuren Softwarelizenzen zu ermöglichen, wird dabei auf die Verwendung des Open-Source-Simulationstools OpenFOAM 9 gesetzt. Durch die Kombination aus automatisierter Geometrieerstellung, Vernetzung und einer vordefinierten, parametrisierten Simulation wird es auch Anwendern ohne einschlägige Simulationsexpertise ermöglicht, CFD-basierte Strömungsanalysen durchzuführen. Der Webservice unterstützt somit eine frühzeitige Integration von Simulation in den Entwicklungsprozess.

Keywords: CFD, OpenFOAM, SaaS, Webservice

Abstract

Numerical simulations are becoming increasingly important in industrial product development. Up to now, their application has traditionally required extensive expertise and specialized software and hardware infrastructure. This is a particular challenge for small and medium-sized enterprises (SMEs). This article presents a cloud-based web service using the example of a flow simulation (CFD simulation) for refrigerated display cases. It was developed to enable simulations to be provided in a cloud environment via a web interface. The web service was applied and tested in the development and optimization process of refrigerated display cases by engineers who are not proven CFD simulation experts. To make this possible without purchasing expensive software licenses, the open-source simulation tool OpenFOAM 9 is used. The combination of automated geometry creation, meshing, and a predefined, parameterized simulation enables users without relevant simulation expertise to perform CFD-based flow analyses. The web service thereby supports the early integration of simulation into the development process.

Keywords: CFD, OpenFOAM, SaaS, Webservice

Einleitung

In der Entwicklung von Kühlmöbeln ist es unumgänglich, im Designprozess darauf zu achten, dass die Konstruktion eine möglichst homogene Temperaturverteilung im Innenraum und einen niedrigen

Energiebedarf erzielt. Gerade im Fall von offenen Kühlmöbeln ist das Zusammenspiel zwischen der Luftströmung und der Wärmeübertragung innerhalb und außerhalb des Warenraums sehr komplex, da ein kühler Luftvorhang den Warenraum von der warmen Umgebung trennen muss.

Konventionell werden Kühlmöbeln in einem iterativen Prozess zwischen Konstruktion, Prototypenbau und Produktion entwickelt. Damit einher gehen hohe Prototypen- und Versuchsaufwände inklusive umfassender normgerechter Messungen und Prüfungen. Durch den Einsatz von digitalen Entwicklungsmethoden basierend auf numerischen Simulationen kann dieser Prozess maßgeblich unterstützt werden. Weiters ergeben sich dadurch neue Möglichkeiten, die Produkte weiter zielgerichtet zu optimieren, sodass im späteren Betrieb der Kühlmöbel Energie eingespart und so ein wichtiger Beitrag zur Nachhaltigkeit geleistet werden kann. Zusätzlich können Ressourcen, Aufwände, Zeit und Kosten in der Produktentwicklung reduziert werden.

Die Schwierigkeit besteht dabei darin, dass die nötige Strömungssimulationsmethode (CFD-Simulation) einerseits sehr komplex und daher für Nicht-Expert:innen schwer zugänglich ist, andererseits mit hohem Aufwand und in der Regel auch mit hohen Kosten für Software und Hardware verbunden ist. Um diese digitalen Simulationswerkzeuge für einen konkreten Anwendungsbereich auch für Nutzer:innen ohne umfassende Simulationsexpertise zugänglich zu machen, ist ein robustes, parametrisiertes und automatisiertes Simulationsmodell erforderlich. Dieses ist in einer Cloud Umgebung implementiert und wird den Anwender:innen über ein gut zugängliches und intuitives Web-Interface zur Verfügung gestellt. Um eine gute Nutzbarkeit und Funktionalität für den Anwender:innen zu schaffen, muss man einen Kompromiss zwischen Allgemeingültigkeit, Detailgrad und Genauigkeit eingehen. Außerdem muss im Hinblick auf Genauigkeit auch der Rechenaufwand der Modelle berücksichtigt werden. Weiters ist es unumgänglich, die Robustheit des Modellierungs- und Simulationsworkflows zu gewährleisten, um eine Implementierung des Webservices zu ermöglichen, die auch für Nicht-Expert:innen zuverlässig nutzbar ist.

Vor diesem Hintergrund richtet sich der Fokus in der Forschung und Entwicklung zunehmend auf automatisierte, parametrisierbare und cloudbasierte Simulationslösungen, welche darauf abzielen, Simulationsanwendungen einfacher zugänglich zu machen. Dadurch verlagert sich einerseits die Rechenlast auf die Cloud und andererseits ermöglichen automatisierte Vernetzungsalgorithmen, automatisierte Simulationsworkflows und webbasierte Benutzeroberflächen einen niederschwelligen Zugang zu Simulation. Damit kommt es nach und nach zu einem Paradigmenwechsel von der durch Expert:innen durchgeführten simulationsgetriebenen Produktentwicklung hin zur zugänglicheren, design-integrierten Simulationsunterstützung auch für Nicht-Simulationsexpert:innen.

Dieses Phänomen sieht man auch bei großen Softwareherstellern, die zunehmend auf Cloud- und webbasierte Services mit teilautomatisierten Workflows setzen. So gibt es mittlerweile zum Beispiel von ANSYS [2], Siemens [3] oder AVL [4] Angebote, die “Simulation as a Service” (SaaS) über Web- und Cloud-basierte Anwendungen anbieten. Andere Softwarehersteller wie SimScale [5] setzen darauf, Simulationen nicht nur via Web-Interface zu bedienen und auf der Cloud zu rechnen, sondern auch AI-gestützte automatisierte Workflows zur Unterstützung der Ingenieur:innen von der Erstellung der Simulation bis zur Analyse der Ergebnisse einzusetzen.

Ziel dieses Beitrags ist es, am Beispiel der Entwicklung von Kühlmöbeln einen Funktionsprototyp für einen Webservice vorzustellen, der auch von Nicht-Simulationsexpert:innen im Produktentwicklungs- und Optimierungsprozess genutzt werden kann. Um das auch ohne die Anschaffung von teuren Softwarelizenzen zu ermöglichen, werden dabei Open-Source-Simulationstools verwendet.

Lösungsweg

CFD-Modell

Um ein effizientes Testen des Funktionsprototyps zu ermöglichen, wurde ein stark vereinfachtes Testsetting für das CFD-Simulationsmodell verwendet. So kann die Funktionalität des Prototyps mit

einem vertretbaren Rechenaufwand anhand verschiedenster Szenarien in einer angemessenen Zeit getestet werden. Für die CFD-Simulation wird die Open-Source-Software OpenFOAM 9 [1] eingesetzt. In OpenFOAM 9 sind die Vernetzer BlockMesh und SnappyHexMesh integriert, die zur Vernetzung der Geometrie verwendet werden. Sowohl die Vernetzung als auch die Simulation können im Funktionsprototyp entweder auf einem CPU-Kern im Modus „*Single*“ oder auf 4 CPU-Kernen im Modus „*Parallel*“ durchgeführt werden.

Geometrie

Die Geometrie für das im Funktionsprototyp verwendete Testsetting wurde an die Geometrie eines in einem Testraum stehenden offenen Kühlmöbels angelehnt. Zur Vereinfachung wurden unter anderem die Regalböden des Kühlmöbels in der Geometrie weggelassen und die Struktur der Rückwand vereinfacht. In Abbildung 1 ist die vereinfachte Geometrie des Kühlmöbels im Testraum dargestellt. Die Kühlmöbelgeometrie ist im Funktionsprototyp als *.stl-File hinterlegt. Im Funktionsprototyp kann nur die Gesamtgröße des Kühlmöbels variiert werden. Dabei wird das stl-File, welches die Geometrie des Kühlmöbels beinhaltet, in allen drei Koordinatenrichtungen um einen vorgegebenen Faktor skaliert, der als Parameter im Web-Interface eingestellt werden kann.

Die Größe des quaderförmigen Testraumes kann vom Nutzer des Frameworks durch Festlegen der Eckpunkte (*X Min*, *Y Min*, *Z Min*) und (*X Max*, *Y Max*, *Z Max*) entlang der Hauptdiagonale des Quaders frei gewählt werden.

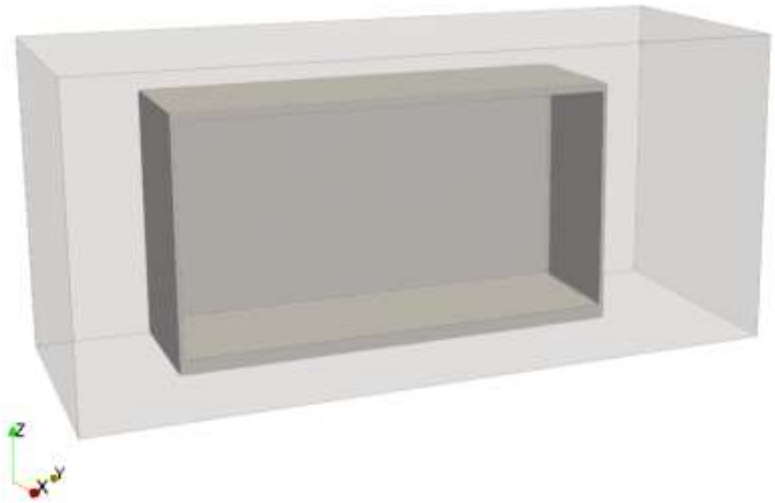

Abbildung 1. Im Funktionsprototyp verwendete vereinfachte Geometrie.

Vernetzung

Mit BlockMesh wird ein quaderförmiges Hintergrundnetz für den Testraum erstellt. Die Auflösung dieses Netzes in jede Koordinatenrichtung kann vom Nutzer mithilfe der Parameter *X Discr*, *Y Discr* und *Z Discr* vorgegeben werden. Mit diesen Parametern wird festgelegt, wie viele Zellen beim Vernetzen des Quaders in der jeweiligen Raumrichtung erzeugt werden.

Im Anschluss an die Erstellung des Hintergrundnetzes wird SnappyHexMesh verwendet, um aus dem Hintergrundnetz des Testraumes die Geometrie des Kühlmöbels auszuschneiden. Dabei wird das Netz annähernd an die Oberfläche der *.stl-Geometrie angepasst, indem das Hintergrundnetz durch Halbierung der Zellen in der Nähe der Oberfläche der *.stl-Geometrie iterativ verfeinert wird. Im Anschluss an diese Verfeinerung werden die Zellen entfernt, welche außerhalb des Fluidvolumens liegen. Dabei entsteht ein Netz, das die Flächen der *.stl-Geometrie sehr gestuft abbildet, da die Zellen entlang der Kanten quaderförmig sind. Deshalb werden danach die Zellknotenpunkte, die entlang der Oberfläche der *.stl-Geometrie liegen, auf die Oberfläche verschoben, um die gestuften Flächen zu glätten und besser an die Oberfläche anzupassen. Um auch große Geschwindigkeitsgradienten nahe den

Wänden in der Simulation ausreichend auflösen zu können, werden an den Wandflächen zusätzliche Schichten mit hexaedrischen Zellen eingefügt, sogenannte Inflationsschichten. Direkt an der Wand sind diese eingefügten Zellen sehr dünn, die Schichtdicke nimmt dann von den Wänden ins Innere des Simulationsbereichs exponentiell zu [6]. Die Einstellungen für SnappyHexMesh sind im Funktionsprototyp fest vorgegeben und können vom User nicht verändert werden, da dies Expertise in der Vernetzung von Geometrien für die CFD-Simulation zur gezielten Beeinflussung und Optimierung der Netzqualität erfordern würde.

Simulationsmodell

Die Simulation in diesem Funktionsprototyp dient zum Testen der Funktionalität des Web-Interfaces und des Konzepts „Simulation-on-demand“ für Nicht-Simulationsexperten. Um das effizient zu ermöglichen, ist die Komplexität der Simulation stark reduziert. Es wird eine stationäre, isotherme und inkompressible Strömung angenommen. Der Nutzer des Funktionsprototyps kann als Randbedingungen den Druck am Auslass aus dem Simulationsbereich bzw. Umgebungsdruck *Ambient Pressure* und die vektorielle Strömungsgeschwindigkeit am Einlass in den Simulationsbereich *Inlet Velocity* vorgeben. Außerdem kann der Nutzer mit dem Parameter *End time* einstellen, wie viele Iterationen in dieser stationären Simulation gerechnet werden sollen. Die restlichen Randbedingungen, genauso wie die Dichte und Viskosität der Luft und alle anderen numerischen Einstellungen sind im Funktionsprototyp fest vorgegeben.

Wählt man für die Durchführung der Simulation den Modus „Parallel“ aus, dann wird die Simulation im Funktionsprototyp auf 4 CPU-Kernen parallel ausgeführt. Bei Netzen mit vielen Zellen kann die Simulation dadurch beschleunigt werden. Hat das Netz nur sehr wenige Zellen, dann kann eine parallelisierte Simulation kontraproduktiv sein, weil der Speed-Up durch die Parallelisierung geringer ist als der Mehraufwand durch die zusätzliche Kommunikation zwischen den CPU-Kernen. Als Richtwert gilt in der vorliegenden Anwendung, dass eine Parallelisierung sinnvoll wird, wenn pro Prozessorkern mehr als 25 000 Zellen anfallen.

Darstellung der Simulationsergebnisse

Der Funktionsprototyp verfügt über eine automatisierte Darstellung der Simulationsergebnisse. Dabei werden *.vtk-Files erzeugt, welche die Datenbasis bilden, um im Viewer das Netz und Simulationsergebnisse anzeigen zu können. Die Aufbereitung der Ergebnisse wird im Anschluss an die Simulation automatisch ausgelöst. Im Web-Viewer können nur Flächen und kein Volumen dargestellt werden. Deshalb müssen die automatisch mit OpenFOAM erstellten *.vtk-Files zuerst in *.vtp-Files umgewandelt werden, um das Netz und die Simulationsergebnisse im Web-Viewer laden und ansehen zu können. Verfügbar sind hier alle Randflächen im Simulationsmodell so wie Druck bzw. Geschwindigkeit auf ausgewählten, im Simulationsmodell fest hinterlegten Schnittflächen durch die Geometrie.

Bereitstellung als Webservice

Um eine Bereitstellung der Simulation in einer Cloudumgebung via Web-Interface zu ermöglichen, wurde ein Python-Modul entwickelt, über das Nutzer:innen Input-Parameter in der Vernetzung und Simulation setzen können. Erst über ein Application Programming Interface (API) wird das Python-Modul zur Konfiguration der Nutzer:innen-Parameter von externen Diensten aus ansprechbar, sodass insbesondere auch eine Kommunikation über ein Web-Interface möglich wird. Auf diese Weise kann das Simulationsframework als Webservice bereitgestellt werden und in externe Anwendungen oder Systeme integriert werden.

Die API liefert in Zusammenhang mit einem containerisierten OpenFOAM die Grundlage dafür, dass die Simulation dezentral ohne Hardware- und Software-Ressourcen bei den Anwender:innen in einer Cloud-Umgebung durchgeführt werden kann. Durch ein Web-Interface können Nutzer:innen des Webservices über die API die verfügbaren Parameter via Graphical User Interface verändern, die Vernetzung und Simulation starten und die Simulationsergebnisse darstellen.

Der in diesem Beitrag vorgestellte Funktionsprototyp wurde in der Cloud-Umgebung von Amazon Web Services [7] unter Einsatz von Amazon Elastic Container Service (ECS) [8] auf EC2 [9] implementiert und den Anwender:innen zugänglich gemacht.

Ergebnis

In Abbildung 2 sind die Eingabemasken für die Geometrieerstellung, Vernetzung und Simulation dargestellt. In Abbildung 3 ist als exemplarisches Beispiel für die Darstellung der Simulationsergebnisse ein Konturplot des Geschwindigkeitsbetrags auf der Schnittfläche normal zur y-Richtung durch die Mitte der Geometrie abgebildet.

Durch die Automatisierung der Vernetzung und Simulationserstellung mittels API und Python-Modul ist für den Endnutzer des Webservice kein tiefes Fach- und Numerik-Wissen nötig, um die Simulation als Werkzeug in der Produktentwicklung zu nutzen. Da für die Durchführung der Simulation eine Cloud-Umgebung genutzt wird, ist die Simulation jederzeit und ortsunabhängig einsetzbar. Die Anwender müssen sich keine spezielle, leistungsstarke Hardware anschaffen und keine Software installieren, da die Simulation als Webservice konzipiert ist und über ein Web-Interface im Browser bedient werden kann.

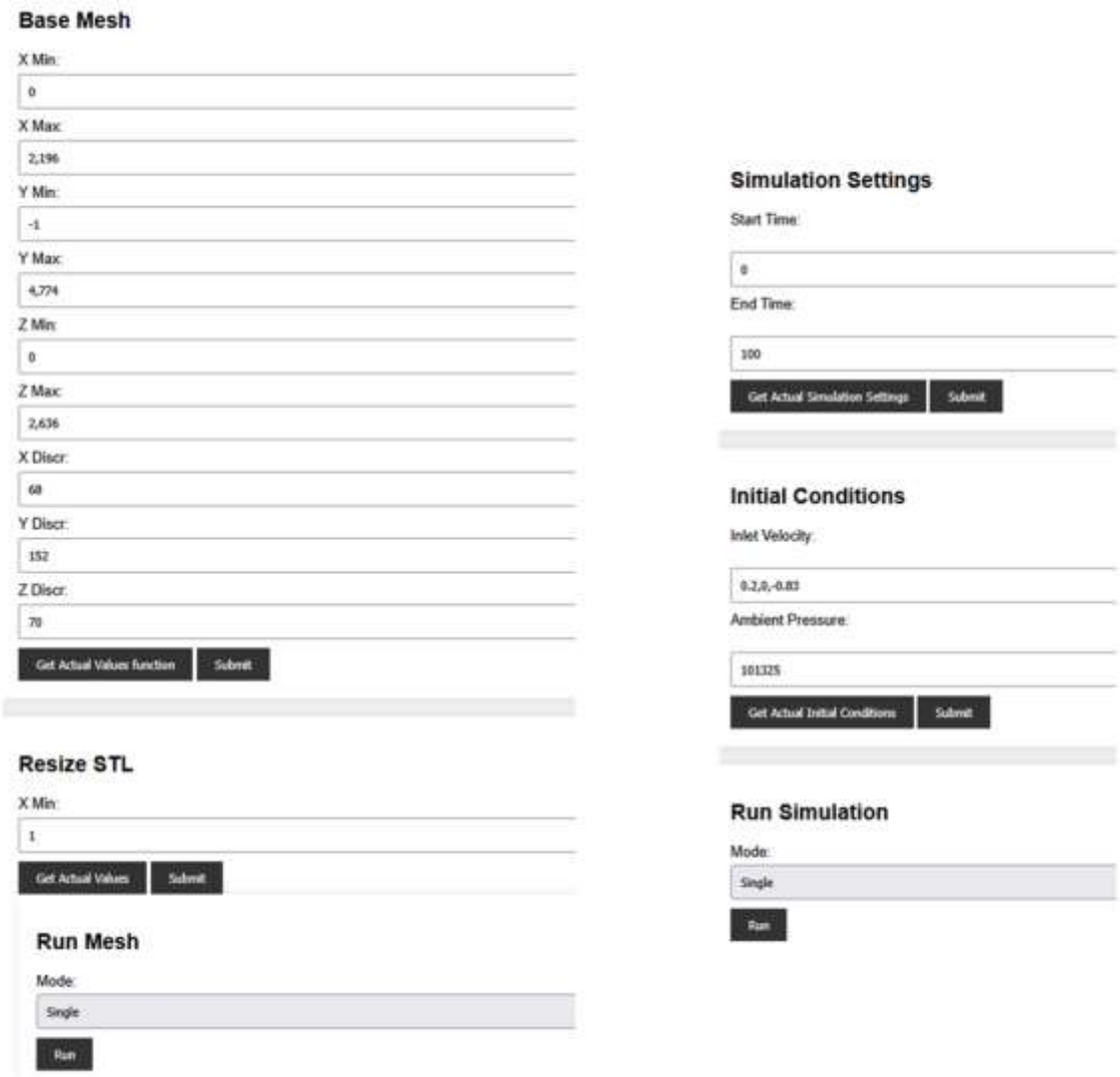

Abbildung 2. Eingabemaske des Web-Interfaces für die automatisierte Vernetzung (links) und für die automatisierte Simulation (rechts).

Zukünftig ist es auch denkbar, den Webservice um einen Optimierer zu ergänzen und so zum Beispiel automatisierte Geometrieoptimierungen durchzuführen. Dazu wäre es zusätzlich noch nötig, eine Schnittstelle zu einer Open-Source CAD-Software einzubauen, damit die Geometrie parametrisiert erzeugt und somit auch unkompliziert und automatisiert verändert werden kann. So könnte eine numerisch unterstützte Optimierung auch für Nicht-Simulationsexpert:innen zugänglich gemacht werden und eine wesentliche Unterstützung im Designprozess neuer Produkte bieten.

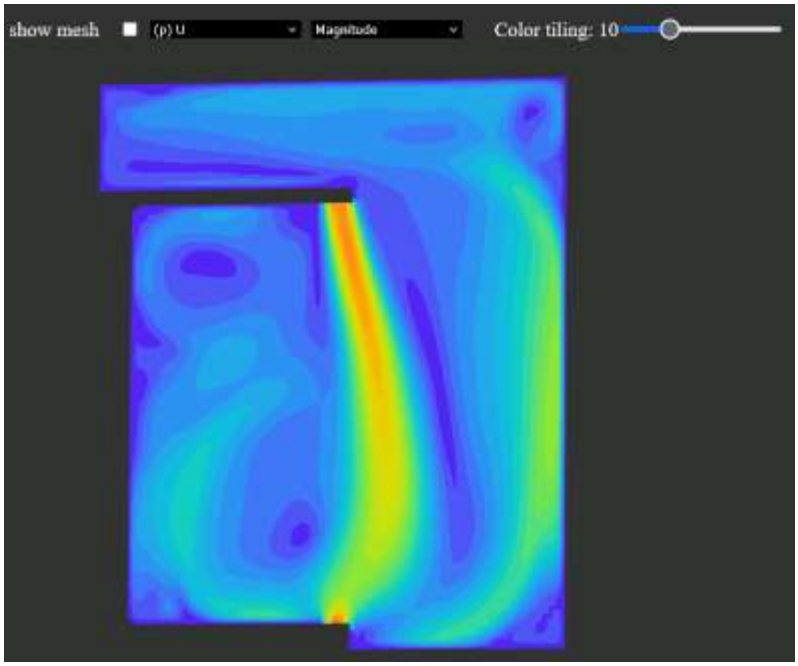

Abbildung 3. Exemplarischer Konturplot des Geschwindigkeitsbetrags.

Des Weiteren kann das entwickelte Framework so angepasst werden, dass es auch für Simulationen aus anderen physikalischen Domänen, z.B. wie FE-Simulationen, sowohl mit Open-Source Tools als auch mit kommerzieller Software, angewendet werden kann. Der Webservice ist damit sehr vielfältig anwendbar und kann zukünftig auch in anderen Branchen und Unternehmen an die individuellen Bedürfnisse angepasst und nutzbringend eingesetzt werden.

Zusammenfassung

Es wurde ein Funktionsprototyp zur automatisierten Remote-Berechnung eines geeignet vereinfachten CFD-Simulationsmodells am Beispiel der Kühlmöbelentwicklung erstellt und getestet. Dabei wurde gezeigt, dass ein guter Kompromiss zwischen Nutzbarkeit und Funktionalität bzw. Detailgrad für die Anwendungsbereiche und Anforderungen gefunden werden kann. Im Zuge des Praxistests des Webservice durch Nicht-Simulationsexpert:innen wurden die Zugänglichkeit und einfache Anwendbarkeit als sehr positiv bewertet. Weiters wurde hervorgehoben, dass der Webservice eine gute Grundlage für die niederschwellige Integration von CFD-Simulation im Entwicklungsprozess von Kühlmöbeln darstellt. Es wurden auch einige Punkte identifiziert, bei denen es noch Erweiterungsbedarf und Verbesserungspotential sowohl auf Ebene der GUI als auch hinsichtlich der technischen Funktionalität gibt. Die isotherme Simulation des Funktionsprototypen kann zu einer thermischen Simulation erweitert werden. Auch die Integration einer Schnittstelle zu einer Open-Source CAD-Software ist wünschenswert, damit unterschiedliche Kühlmöbel-Geometrien parametrisiert erzeugt und unkompliziert variiert werden können.

Des Weiteren kann man das Framework durch diverse Anpassungen auch für Simulationen aus anderen physikalischen Domänen wie z.B. FE-Simulationen verwenden. In das Framework kann neben Open-Source Tools auch kommerzielle Software integriert werden.

Auf diese Weise kann ein Webservice für numerische Simulation aufgebaut werden, der auf niederschwellige Art und Weise zur Unterstützung im Design- und Entwicklungsprozess von Produkten eingesetzt werden kann. Dadurch lassen sich neue Potentiale erschließen, um Zeit und Kosten einzusparen.

Literatur

[1] H. Weller, C. Greenshields, und W. Bainbridge, *OpenFOAM 9*. (20. Juli 2021). C++. The OpenFOAM Foundation Ltd, London. [Online]. Verfügbar unter: https://openfoam.org/release/9/

[2] „Simulation as a Service: Ansys Lizenzen & Hardware aus der Cloud“. Zugegriffen: 23. Juli 2025. [Online]. Verfügbar unter: https://www.cadfem.net/at/de/unsere-loesungen/simulation-as-a-service.html

[3] „Simcenter X - HPC and remote desktop“, Siemens Digital Industries Software. Zugegriffen: 23. Juli 2025. [Online]. Verfügbar unter: https://plm.sw.siemens.com/de-DE/simcenter/simcenter-x/hpc/

[4] „Simulation as a Service | AVL“. Zugegriffen: 23. Juli 2025. [Online]. Verfügbar unter: https://www.avl.com/en/simulation-solutions/simulation-a-service

[5] „Simulation Software | Engineering in the Cloud“, SimScale. Zugegriffen: 23. Juli 2025. [Online]. Verfügbar unter: https://www.simscale.com/

[6] C. Greenshields, „OpenFOAM v9 User Guide“, CFD Direct. Zugegriffen: 22. Juli 2025. [Online]. Verfügbar unter: https://doc.cfd.direct/openfoam/user-guide-v9/snappyhexmesh/

[7] „Amazon Web Services AWS – Server Hosting & Cloud Services“, Amazon Web Services, Inc. Zugegriffen: 25. Juli 2025. [Online]. Verfügbar unter: https://aws.amazon.com/de/

[8] „AWS | Amazon EC2 Container & Konfigurationsmanagement“, Amazon Web Services, Inc. Zugegriffen: 22. Juli 2025. [Online]. Verfügbar unter: https://aws.amazon.com/de/ecs/

[9] „Amazon EC2 – Cloud-Datenverarbeitungskapazität – AWS“, Amazon Web Services, Inc. Zugegriffen: 22. Juli 2025. [Online]. Verfügbar unter: https://aws.amazon.com/de/ec2/

Intelligente Chatbots zur Abfrage von Unternehmenswissen - ALLEX

Pichlmann Lukas
LKR Leichtmetallkompetenzzentrum
Ranshofen GmbH
Lukas.Pichlmann@ait.ac.at

Kurzzusammenfassung

Die Implementierung eines intelligenten Chatbots Advanced Language Learning for Engineering Excellence (ALLEX) zur Abfrage von Unternehmenswissen stellt eine nutzerfreundliche Lösung zur Verbesserung der Informationszugänglichkeit dar. Dieser Beitrag beschreibt die Entwicklung eines Chatbots basierend auf Retrieval Augmented Generation, welcher Nutzeranfragen zu firmeninternen Dokumenten mittels lokaler Large Language Models (LLMs) beantwortet. Die Lösung nutzt Open Source Komponenten und kombiniert semantische Suche mit lokalen LLMs. Das System ermöglicht es Mitarbeitern und Mitarbeiterinnen, effizient auf historische Daten und Dokumente zuzugreifen und dabei von intelligenter Kontextualisierung zu profitieren.

Keywords: RAG, Chatbot, Large Language Models

Abstract

The implementation of an intelligent chatbot for querying company knowledge represents a user-friendly solution for improving information accessibility. This contribution describes the development of a chatbot based on Retrieval Augmented Generation, which answers user queries about internal company documents using local Large Language Models (LLMs). The solution utilizes open source components and combines semantic search with local LLMs. The system enables employees to efficiently access historical data and documents while benefiting from intelligent contextualization.

Keywords: RAG, Chatbot, Large Language Models

Einleitung

Unternehmen sammeln über Jahre hinweg umfangreiche Datenbestände an Dokumentationen und Erfahrungswissen. Diese Informationen sind oft schwer zugänglich oder schwer auffindbar und werden deshalb nicht optimal genutzt. Mitarbeiter:innen profitieren von einer nutzerfreundlichen Möglichkeit, um auf dokumentierte Fälle und Prozesse im Unternehmen zuzugreifen, ähnliche Situationen zu identifizieren und aus historischen Lösungsansätzen zu lernen.

Retrieval Augmented Generation (RAG) kombiniert die Stärken von Informationsabruf und Textgenerierung und ermöglicht es, spezifisches Domänenwissen in Large Language Models (LLM) zu integrieren, ohne aufwendige Finetuning-Prozesse durchführen zu müssen.

Der hier implementierte Ansatz ist branchenunabhängig anwendbar und lässt sich auf verschiedene Unternehmenskontexte übertragen, um die Abfrage von internem Firmenwissen zu optimieren und Mitarbeitern präzise, kontextbezogene Antworten auf ihre Fragen zu liefern.

Um ein RAG-System zu implementieren, werden verschiedene und komplexe Technologien benötigt, deren Neuentwicklung den Rahmen dieses Beitrags bei weitem überschreiten würde. Daher wurde auf bewährte Open-Source (OS) Komponenten gesetzt, die von der kontinuierlichen Weiterentwicklung und dem kollektiven Wissen der Community profitieren und gleichzeitig eine vergleichsweise hohe Qualität und Stabilität gewährleisten. Zudem wurde darauf geachtet, kleine und ressourcenschonende LLMs zu verwenden, um Hardware-Anforderungen zu minimieren und das System mit möglichst geringen Investitionen zugänglich zu machen.

Problemlösungsweg / Vorgehen

Systemarchitektur

Für die Umsetzung von ALLEX wurden drei verschiedene Open Source Projekte ausgewählt:

- Hosting von lokalen LLMs: Ollama
- Hosting von Embedding und Reraking Modellen für RAG: XInference
- Plattform mit RAG-Unterstützung, Web-Interface sowie (Low-Code) Funktionalität zur Programmierung der Chatbot Logik: Dify

Alle Komponenten können lokal, offline und containerisiert mittels Docker gehostet werden und bieten eine freizügige Lizenz, was eine der wichtigsten Bedingungen in der Technologieauswahl darstellte.

Ollama

Ollama ist ein Tool, das entwickelt wurde, um die Bereitstellung und die lokale Anwendung von Large Language Models (LLMs) zu erleichtern [1]. Ollama bietet zudem eine umfassende Modell-Bibliothek mit einer breiten Auswahl an State Of The Art Open Weight LLMs, die mit geringem Aufwand verwendet werden können.

Ollama wurde ausgewählt, da es ein beliebtes Tool mit einer sehr großen Open-Source-Community ist, was einen signifikanten Grad an Stabilität gewährleistet. Darüber hinaus zeichnet es sich durch eine einfache Installation und Benutzerfreundlichkeit aus, ermöglicht die lokale Ausführung verschiedener LLMs ohne komplexe Konfiguration und bietet eine aktive Entwicklergemeinschaft, die regelmäßig Updates und Verbesserungen bereitstellt.

XInference und semantische Suche

Ähnlich wie Ollama stellt XInference eine Möglichkeit für das lokale Hosting von LLMs, aber auch anderen KI-Modellen, dar [2]. Da zum Zeitpunkt der Entwicklung von ALLEX Ollama nicht die Möglichkeit bot, sogenannte Reranking Modelle zu hosten, wurde XInference als weitere Komponente von ALLEX ausgewählt. In ALLEX wird XInference dazu verwendet, die nötigen Modelle für die spätere semantische Suche bereitzustellen.

Um semantische Suche ausführen zu können, werden sogenannte Embedding Modelle benötigt. Diese sind darauf spezialisiert, natürlichen Text und dessen Bestandteile als (hochdimensionale) numerische Vektoren darzustellen, die den semantischen Kontext bestimmter Textabschnitte repräsentieren. Damit können Informationen, wie beispielsweise firmeninterne Dokumente und der darin enthaltene Text, durch maschinenlesbare Zahlenwerte abgebildet werden. Diese numerischen Daten können dann in eine Vektordatenbank gespeichert werden, wodurch sie für die semantische Suche bereitstehen. Mittels solcher Embedding Modelle werden später auch eingehende Nutzeranfragen in numerische Vektoren umgewandelt. Mittels dieser Vektoren und einer vordefinierten Distanzmetrik (beispielsweise Kosinus-Ähnlichkeit) kann dann die Vektordatenbank nach ähnlichen Einträgen (also nach Vektoren mit geringer Distanz zum aus der Anfrage resultierenden Vektor) durchsucht werden.

Einträge aus der Datenbank werden mit Hilfe der Distanzmetrik nach Relevanz gereiht. Um die Präzision dieser Reihung zu verbessern, können optional sogenannte Reranking Modelle verwendet werden. Diese Modelle bewerten die durch die initiale Vektorsuche gefundenen Kandidaten nochmals detaillierter und ordnen sie basierend auf einer (verbesserten) Einschätzung der Relevanz zur Nutzeranfrage neu an. Reranking-Modelle arbeiten typischerweise mit sogenannten Cross-Attention-Mechanismen, welche die semantische Beziehung zwischen Anfrage und Dokumenten präziser erfassen können als die reine Vektordistanz.

XInference bietet eine umfassende Auswahl an Open Weights Embedding und Reranking Modellen, welche in ALLEX für die Umsetzung von Retrieval Augmented Generation verwendet werden.

Dify

Ollama und Xinference gewährleisten das lokale Deployment aller nötigen KI-Modelle für das System. Die Hauptarbeit dieses Beitrags steckt jedoch in der Logik und der Programmierung des Chatbots, welcher eine Kombination von LLM, Embedding und Reranking Modellen, Dokumentendatenbank und Programmlogik darstellt. Zum Erstellen dieser Programmlogik sowie zur Konfiguration der Datenbank wurde Dify ausgewählt [3].

Dify ist ein modulares OS-Projekt, welches sich auf die Entwicklung von LLM-Anwendungen spezialisiert. Es bietet eine einfache Einbindung verschiedenster Hosting Services, darunter auch Ollama und XInference, und ermöglicht die einfache Konfiguration einer Vektordatenbank sowie die Entwicklung komplexer LLM-Workflows mit Hilfe einer Low-Code Plattform.

Abbildung 1 zeigt ein minimales Beispiel für einen einfachen Chatflow in der Low-Code Umgebung von Dify, wo zwischen Nutzeranfragen (Start-Node) und zurückgegebener Antwort (Answer-Node) komplexe Logik ausprogrammiert werden kann, entweder mit Hilfe der in Dify vorkonfigurierten Nodes und Bausteine oder mit eigenem Quellcode, welcher in sogenannten Code-Nodes unter Verwendung der Programmiersprache Python ausgeführt werden kann.

Abbildung 1. Beispiel eines einfachen Ablaufs in der Low-Code Umgebung von Dify, bestehend aus einer Start-Node, einem LLM-Node und einer Antwort-Node für einfach Chat-Interaktion mit einem LLM

Durch diese Chatflows können komplexe Workflows abgebildet werden, welche dann Nutzern über ein Chat-Interface zur Verfügung gestellt werden. Zudem bietet Dify die Möglichkeit, verwendete externe Inhalte, die via RAG in der Antwort verwendet wurden, dem Nutzer am Ende der Antwort mit Hilfe von Zitaten anzuzeigen. Diese Funktion ermöglicht es, die Richtigkeit der erzeugten Antworten zu kontrollieren und so die Transparenz des Chatbots zu verbessern.
Dify bietet zudem eine benutzerfreundliche Datenbank-Integration zur Verwendung von RAG, wobei standardmäßig eine Weaviate-Vektordatenbank verwendet wird.

Programmlogik

Abbildung 2 zeigt den im Laufe des Projektes entwickelten Chatflow, der mit Hilfe von Dify erstellt wurde (oben), sowie eine schematische Darstellung der wichtigsten darin enthaltenen Funktionen (unten).

Für die Erstellung dieses Chatflows wurden nicht nur die standardmäßig in Dify enthaltenen Bausteine verwendet, sondern auch selbst entwickelte Nodes, die im Quellcode des Dify Projekts hinzugefügt und programmiert werden mussten: sogenannte “Context”-Nodes. Diese erlauben es, den zur Verfügung stehenden Kontext jedes LLM-Agenten (beziehungsweise jeder LLM-Node) im Chatflow dynamisch anzupassen und während der Ausführung des Chatflows je nach Situation und Entscheidungspfad zu modifizieren. Dies ermöglicht es beispielsweise, nach Generierung einer Antwort auf eine Anfrage die Dokumenteninhalte, die dem LLM für die Beantwortung mittels RAG zur Verfügung gestellt wurden, wieder aus dem Kontext zu entfernen. Dadurch können Token im Kontext-Fenster gespart werden, was

einem LLM-Agenten ein längeres “Gedächtnis” (sprich: mehr vergangene Chat Nachrichten im Kontext) ermöglicht.

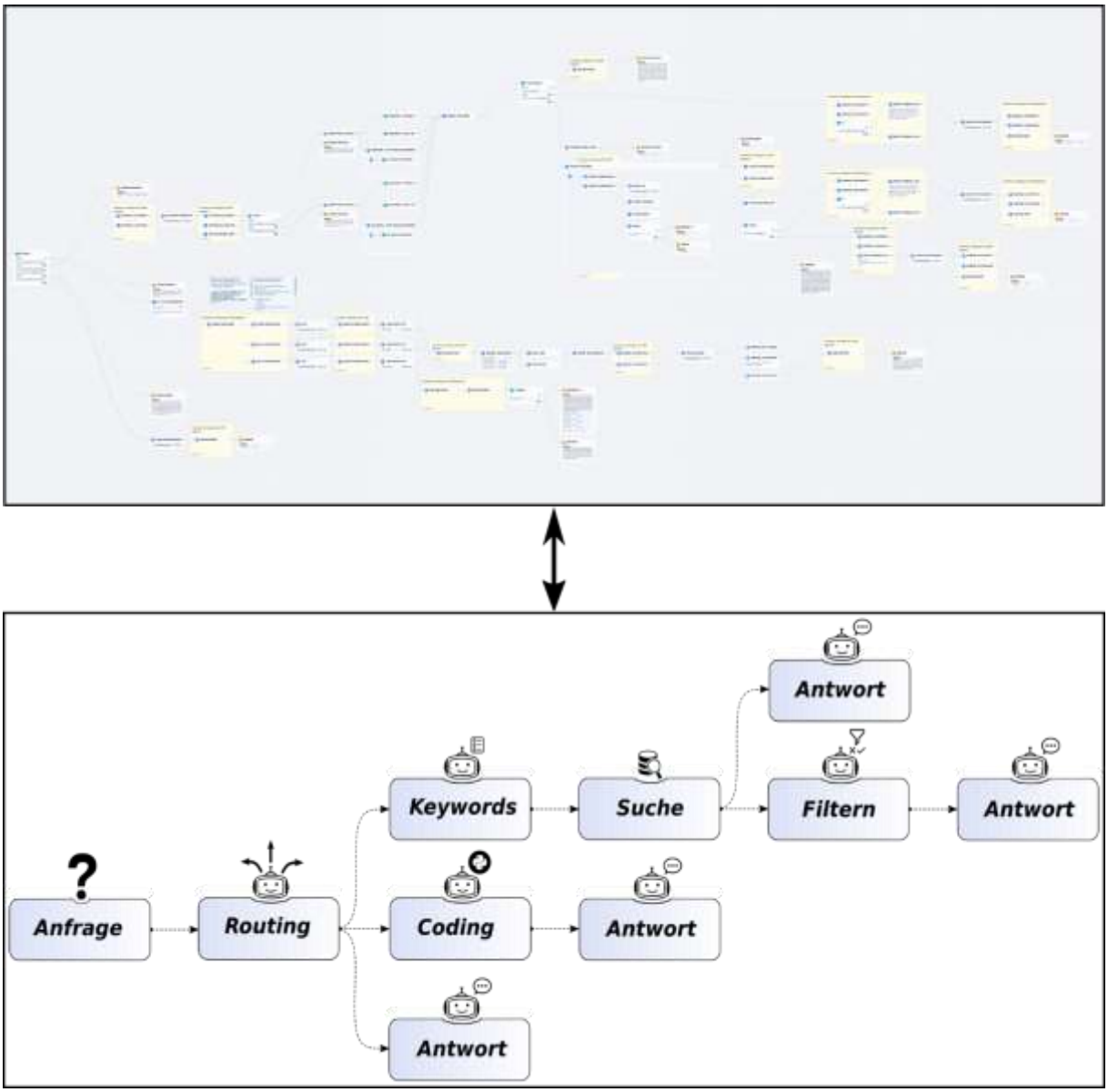

Abbildung 2. Screenshot des erstellten Workflows in Dify (oben) und entsprechende schematische Darstellung der programmierten Logik (unten)

Der Programmablauf besteht aus verschiedenen LLM-Agenten mit unterschiedlichen Aufgaben. Für jeden der involvierten LLM-Agenten wurde ein System-Prompt erstellt, welches die jeweilige Aufgabe ausführlich mit einer Vielzahl an Beispielen erklärt. Jeder LLM-Agent hat zudem eine eigene Chat-Historie, die dynamisch während des Ausführens angepasst wird. Insgesamt werden 5 verschiedene Arten von LLM-Agenten verwendet:

- Der “Routing”-Agent entscheidet, welcher der 3 möglichen Pfade in Abbildung 2 für die aktuelle Anfrage gewählt werden soll und wie diese beantwortet werden soll:
    - Durch direkte Beantwortung ohne weiteren Kontext (unterer Pfad)
    - Mit Hilfe eines Coding Agenten, der Zugriff auf eine strukturierte Tabelle aus historischen Problemfällen hat und so einfache programmatische Analysen ausführen kann (mittlerer Pfad)
    - Mittels RAG, wobei die Anfrage durch ein LLM mit Hilfe von Informationen aus der Datenbank beantwortet wird (oberer Pfad)

- Der “Antwort”-Agent ist dafür zuständig, die finale Antwort zu generieren, die dem Nutzer weitergegeben wird. Dieser wird in jedem der 3 Pfade verwendet.
- Der “Coding”-Agent ist dafür zuständig, Python-Code für einfach programmatische Analysen der Tabelle aus historischen Problemfällen zu generieren.
- Der “Keyword”-Agent ist dafür zuständig, aus der Nutzeranfrage verschiedene Keywords zu generieren, die Zusammen mit der Anfrage für die semantische Suche verwendet werden.
- Der “Filter”-Agent wird optional (je nach Nutzer-Konfiguration) nach der semantischen Suche verwendet, um für jeden der ausgewählten Einträge zu entscheiden, ob der jeweilige Eintrag tatsächlich bei der Beantwortung der Anfrage hilft, oder aus dem Kontext entfernt werden kann. Dies hilft dabei den begrenzten Kontext des LLMs nicht mit irrelevanten Informationen zu befüllen.

Hier ein Beispiel für den so entstehenden Workflow: Der Nutzer stellt im Chat-Interface die Frage “Gabe es schon einmal Probleme mit falsch bemaßten Bauteilen und was wurde unternommen?”. Der “Routing”-Agent entscheidet, dass für diese Nutzeranfrage semantische Suche benötigt wird (RAG). Der “Keyword”-Agent bekommt die aktuelle Anfrage (sowie eine begrenzte Chat-Historie) und generiert eine Handvoll Keywords und Key-Phrases. Mit Hilfe der originalen Anfrage sowie den generierten Keywords wird die Vektordatenbank nach relevanten Einträgen durchsucht, und die Top-N Einträge, deren Relevanz-Score über einem gewissen Schwellenwert liegt, werden (nach optionalem Aussondern irrelevanter Einträge durch den “Filter”-Agenten) dem “Antwort”-Agenten zur Generierung eine Nachricht zur Verfügung gestellt, welche dann inklusive verwendeter Quellen dem Nutzer ausgegeben wird.

Wird eine Anfrage gestellt, die einer statistischen Analyse oder Auswertung der zur Verfügung stehenden strukturierten Tabelle aus vergangenen Problemfällen im Unternehmen bedarf (beispielsweise “Wie viele Problemberichte gabs im Jahr X und was waren die durchschnittlich verursachten Kosten?”), so sollte der “Routing”-Agent den Pfad zur programmatischen Analyse auswählen. Der dann durch den “Coding”-Agenten generierte Code wird in einer Python-Sandbox ausgeführt, und wenn erfolgreich wird das Resultat dem “Antwort”-Agenten übergeben, um eine Antwort gestützt durch die Auswertung zu generieren. Ist das Ausführen nicht erfolgreich, so wird dem “Coding”-Agenten ermöglicht, den Programmcode zu korrigieren. Dazu wird die beim Ausführen entstandene Fehlermeldung gespeichert und dem Agenten für die Korrektur zur Verfügung gestellt. Der Agent hat maximal 3 Versuche, um validen Python Code zu generieren, bevor der Chatflow abgebrochen wird und dem Nutzer eine entsprechende Fehlermeldung angezeigt wird.

Alle der genannten Agenten verwenden dasselbe zugrunde liegende LLM und werden sequenziell ausgeführt, lediglich mit anderem Kontext. So wird sichergestellt, dass zum Betrieb des Chatbots ein einziges im GPU-Speicher befindliches Modell genutzt werden kann, was für geringere Hardwareanforderungen sorgt.

Evaluierung

Um die Qualität der Antworten des Systems zu evaluieren und basierend darauf die Programmlogik, verwendete Modelle, sowie Parameter anzupassen, wurde ein umfassender Fragenkatalog erarbeitet. Darin wurden verschiedenste mögliche Nutzeranfragen durch ALPEX definiert sowie erwartete Antworten und relevante Dokumente, die zur Beantwortung benötigt werden. Damit konnte auf objektive Weise festgehalten werden, wie viele der relevanten Textausschnitte und Dokumente bei der Beantwortung der Anfrage tatsächlich mittels semantischer Suche gefunden wurden. Zudem wurde durch ein unabhängiges (lokales) LLM bewertet, wie gut die Übereinstimmung der Chatbot-Antwort mit der erwarteten Antwort aus dem Fragenkatalog ist (Score von 1 bis 5).
Da Dify das programmatische Ausführen des Chatbots über eine eigene API ermöglicht, konnte dieser Evaluierung-Prozess zur Gänze automatisiert werden.

Dafür wurde ein Python Skript erstellt, welches über die Fragen und erwarteten Antworten/Dokumente im Fragenkatalog iteriert, mittels Dify-API den Chatbot für die jede Frage ausführt, und die resultierenden Antworten inklusive Zitate mit den verwendeten Dokumenten auswertet. Metriken wie die Anzahl der gefundenen relevanten Dokumente und Bewertung der Übereinstimmung (durch ein LLM) mit der vordefinierten erwarteten Antwort wurden dann für jede Frage gespeichert und schlussendlich aggregiert, um die Performance des Systems zu quantifizieren.

Basierend auf diesem Score wurde die Programmlogik überarbeitet, Modelle angepasst, und Parameter in der semantischen Suche optimiert.

Für den erstellten Prototypen wurden schlussendlich folgende Modelle ausgewählt:

- LLM: Mistral Small 3.1 24B (Q6) [4] [5]
- Embedding Modell: Multilingual e5 Large [6]
- Reranking Modell: Jina Reranker v2 [7]

Die genannten Modelle zeigten vielversprechende Performance und stellen zudem eine ressourcenschonende Möglichkeit dar, das System zu betreiben. Mit, für LLM-Anwendungen, vergleichsweise geringem benötigen GPU-VRAM von etwa 30GB kann die Software zum Zeitpunkt des Schreibens schon mit handelsüblicher Consumer-Hardware wie beispielsweise der NVIDIA 5090 performant genutzt werden.

Ergebnis

Das System bietet den Nutzern die Möglichkeit, durch natürliche Sprache mit einem Chatbot zu kommunizieren, welcher dynamisch entscheidet wie und mit Hilfe welcher Werkzeuge die Nutzeranfrage beantwortet werden soll. Dabei können zur Beantwortung der Anfrage Informationen aus einer Datenbank genutzt werden oder auch einfache Programme zur Datenauswertung erstellt und ausgeführt werden, falls nötig.

Der Fokus des Systems liegt darauf, alle verwendeten Komponenten vollständig offline betreiben zu können, um hohe Datensicherheit zu gewährleisten. Zudem wurde darauf geachtet, möglichst kleine und ressourcenschonende LLMs zu verwenden, um das System auch mit vergleichsweise kostengünstiger Hardware betreiben zu können. Die Lösung bietet so einen effizienten Weg, um moderne Technologien wie RAG auch in kleineren Unternehmen zu integrieren, ohne interne Daten an Cloudprovider zu übergeben.

Zusammenfassung

Der vorliegende Beitrag demonstriert die praktische Umsetzung eines lokalen RAG-Systems für die Abfrage von Unternehmenswissen unter Verwendung von ausschließlich open-source Komponenten. Die entwickelte Lösung kombiniert drei etablierte Technologien (Ollama, XInference und Dify) zu einem funktionsfähigen System, das vollständig offline betrieben werden kann.

Mit Hilfe dieser OS-Tools wurde ein Multi-Agent System mit Datenbankzugriff entwickelt, bei dem verschiedene spezialisierte LLM-Agenten, basierend auf demselben Basismodell, sequenziell arbeiten, um Nutzerfragen zu beantworten. Dies ermöglicht eine effiziente Ressourcennutzung bei gleichzeitig hoher Funktionalität.

Mi der Low-Code Funktionalität in Dify konnte ein Workflow erstellt und iterativ angepasst werden, der es ermöglicht, Nutzeranfragen mit Hilfe von semantischer Suche sowie eines autonomen Coding-Agenten gestützt durch Informationen aus der Datenbank zu beantworten. Die automatisierte Evaluierung mittels Fragenkatalog und LLM-basierter Bewertung erwies sich als effektive Methode zur Bewertung der System-Performance und zur Optimierung der System- und Modellparameter.

Die Lösung zeigt, dass bereits mit vergleichsweise bescheidenen Hardware-Anforderungen (ca. 30GB GPU-VRAM) eine LLM-Applikation realisiert werden kann, welche die Auffindbarkeit von relevanten internen Informationen verbessern kann. Ein solches System ist besonders für kleine und mittlere Unternehmen relevant, da sie eine kostengünstige Alternative zu cloudbasierten Enterprise-Lösungen darstellt. Unternehmen können ihre Datenhoheit vollständig bewahren, während sie dennoch von modernen KI-Technologien profitieren. Durch die ausschließliche Verwendung von open-source Komponenten entstehen außerdem keine Lizenzkosten oder Vendor-Lock-in-Effekte.

Der Beitrag liefert somit einen praktikablen Leitfaden für Unternehmen, die ihre internen Wissensbestände durch moderne KI-Technologien besser erschließen möchten, ohne dabei Kompromisse bei Datensicherheit oder Kosteneffizienz eingehen zu müssen.

Literaturverzeichnis

[1] „Ollama," [Online]. Available: https://ollama.com/. [Zugriff am 28 07 2025].

[2] L. Weizheng, X. Lingfeng, Z. Feng, Q. Xuye, und C. Yueguo, "Xinference: Making Large Model Serving Easy". In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Seite 291–300, Miami, Florida, USA. Association for Computational Linguistics.

[3] LangGenius, „Dify - GitHub," [Online]. Available: https://github.com/langgenius/dify. [Zugriff am 28 07 2025].

[4] „Hugging Face - mistralai/Mistral-Small-3.1-24B-Instruct-2503," [Online]. Available: https://huggingface.co/mistralai/Mistral-Small-3.1-24B-Instruct-2503. [Zugriff am 28 07 2025].

[5] „Hugging Face - bartowski/mistralai_Mistral-Small-3.1-24B-Instruct-2503-GGUF," [Online]. Available: https://huggingface.co/bartowski/mistralai_Mistral-Small-3.1-24B-Instruct-2503-GGUF. [Zugriff am 28 07 2025].

[6] „Hugging Face - intfloat/multilingual-e5-large," [Online]. Available: https://huggingface.co/intfloat/multilingual-e5-large. [Zugriff am 28 07 2025].

[7] „Hugging Face - jinaai/jina-reranker-v2-base-multilingual," [Online]. Available: https://huggingface.co/jinaai/jina-reranker-v2-base-multilingual. [Zugriff am 28 07 2025].

Optimierung von Umformprozessen durch KI-basierte Überwachung und Steuerung

Denk Michael
LKR Leichtmetallkompetenzzentrum
Ranshofen GmbH
Michael.Denk@ait.ac.at

Pichlmann Lukas
LKR Leichtmetallkompetenzzentrum
Ranshofen GmbH
Lukas.Pichlmann@ait.ac.at

Kurzzusammenfassung

Dieser Beitrag präsentiert ein System zur Echtzeitüberwachung von Zugproben mittels klassischer Computer-Vision-Methoden als Grundlage für eine adaptive Stromregelung im Umformprozess. Ziel ist die dezidierte Auslösung des elektroplastischen Effekts (EPE) durch eine dynamische Anpassung der Stromimpulse in Abhängigkeit der beobachteten Querschnittsveränderung der Probe. Die Geometrieanalyse erfolgt über eine Kombination aus Canny-Kantendetektion und RANSAC-Linienregression zur robusten Bestimmung des Probenabstands. Die gewonnenen Messwerte und davon abgeleiteten Regelparameter dienen unmittelbar als Eingangsgröße für die Steuerung der Hochstromanlage. Das System ermöglicht damit eine prozessnahe, sensorgestützte Steuerung und stellt eine innovative Lösung für die Echtzeitoptimierung industrieller Umformprozesse dar.

Keywords: Computer Vision, Zugprobe, Echtzeitüberwachung

Abstract

This contribution presents a system for real-time monitoring of tensile specimens using classical computer vision methods as a foundation for adaptive current control in forming processes. The objective is the targeted activation of the electroplastic effect (EPE) through dynamic adjustment of current pulses based on the observed cross-sectional changes of the specimen. The geometry analysis is performed through a combination of Canny edge detection and RANSAC line regression for robust determination of specimen spacing. The acquired measurement values and derived control parameters serve directly as input variables for controlling the high-current system. The system thus enables process-oriented, sensor-supported control and represents an innovative solution for real-time optimization of industrial forming processes.

Keywords: Computer Vision, Tensile Specimen, Real-time Monitoring

Einleitung

Die gezielte Auslösung des elektroplastischen Effekts (EPE) in metallischen Werkstoffen eröffnet neue Möglichkeiten zur Reduktion von Umformkräften und Erhöhung der Duktilität. Nach aktuellem Forschungstand ist der Effekt stark abhängig von der Geometrie, dem Materialzustand und der zeitlichen Ansteuerung des Hochstroms. Eine Echtzeitregelung des Stromprofils auf Basis des mechanischen Ist-Zustands der Probe war bisher nicht realisiert.

Bisherige Verfahren arbeiten mit starren Stromprofilen ohne Rückkopplung auf die tatsächlich auftretende plastische Ausdünnung. Ziel des Beitrags war es, ein System zu entwickeln, das über bildgestützte Echtzeiterfassung der Probengeometrie eine adaptive Steuerung der Stromimpulse erlaubt. Damit wird ein intelligenter Closed-Loop-Regelkreis etabliert, der in industriellen Umformprozessen unmittelbar verwertbar ist.

Ein solches bildbasiertes Regelungsprinzip zeigt dabei Übertragungspotenzial auf verschiedene Umformverfahren und bietet insbesondere für kleine und mittlere Unternehmen eine kosteneffiziente Alternative zu aufwendiger Sensortechnik.

Problemlösungsweg / Vorgehen

Das entwickelte System besteht aus einer (GUI-)Anwendung basierend auf folgenden Hauptkomponenten:

- Datenaufnahme: FastAPI-basierte Schnittstelle für die Kameraintegration
- Kamerakalibrierung: Korrektur von Bildverzerrungen
- Parameteranpassung: Anpassung der wichtigsten Parameter für die Kantenerkennung
- Echtzeit-Auswertung: Echtzeitanalyse inklusive Visualisierung

Der Workflow sieht dabei im Wesentlichen wie folgt aus: Mit einer Kamera werden Bilder von der Zugprobe aufgenommen und periodisch via POST-Request an den definierten Endpunkt des Systems geschickt. Zunächst wird ein Schachbrettmuster in die Anlage zur Kalibrierung eingespannt. Der Nutzer kann dann im Graphical User Interface (GUI) des Systems die Kalibrierungsparameter einstellen, um Linsenverzerrungen im Bereich des Schachbrettmusters zu korrigieren und so die anschließende Kantendetektion zu erleichtern. Danach passt der Bediener im GUI die Parameter zur Kantendetektion an, um eine optimale Einstellung für den jeweiligen Versuch sicherzustellen. Nach Ermittlung dieser Parameter kann im GUI die Live-Visualisierung verwendet werden, um die detektierte Probengeometrie, deren Veränderung sowie die Veränderung der damit verbundenen Regelungsparameter (Spannung) zu verfolgen.

Benutzeroberfläche

Für die praktische Anwendung wurde eine intuitive grafische Benutzeroberfläche mit Python und dem PyQt6-Framework entwickelt [1]. Die Anwendung ist in drei funktionale Bereiche unterteilt, die über separate Registerkarten bzw. Tabs zugänglich sind.

Der Live-Ansicht-Tab ermöglicht die Echtzeitüberwachung des Messprozesses mit kontinuierlicher Anzeige des Kamerabilds und der detektierten Kanten. Der Tab zur Parameter-Einstellung bietet Zugriff auf alle relevanten Algorithmus-Parameter wie Canny-Schwellwerte, RANSAC-Parameter und Krümmungsfilter-Einstellungen, um die verwendeten Methoden dynamisch an neue Umgebungen, Lichtverhältnisse und Proben anzupassen. Zuletzt stellt der Kalibrierungs-Tab eine Oberfläche zur Kamerakalibrierung für Verzerrungskorrektur bereit, um Ungenauigkeiten im Fitting-Prozess zu reduzieren.

Kamerakalibrierung

Für präzise Messungen ist die Korrektur der durch die Kamera bedingten Verzerrungen wichtig. Das System nutzt bewährte OpenCV-Kalibrierungsfunktionen zur automatischen Bestimmung und Korrektur von Verzerrungen der Kameraoptik [2]. Als Kalibrierobjekt wird ein standardisiertes Schachbrettmuster verwendet, dessen geometrische Eigenschaften exakt bekannt sind.

Der Kalibrierungsprozess erfolgt durch die automatische Detektion der Schachbrett-Eckpunkte. Aus diesen Punktkorrespondenzen berechnet der Algorithmus die intrinsischen Kameraparameter sowie die extrinsischen Parameter für jede Aufnahme. Die ermittelten Verzerrungsparameter werden anschließend zur Entzerrung aller Messbilder verwendet, wodurch systematische Messfehler durch optische Verzerrungen eliminiert werden.

Abbildung 1 zeigt, wie so mit Hilfe einer Kalibrierungsprobe die Verzerrung im Bereich der Zugprobe unter Verwendung des GUI korrigiert werden kann.

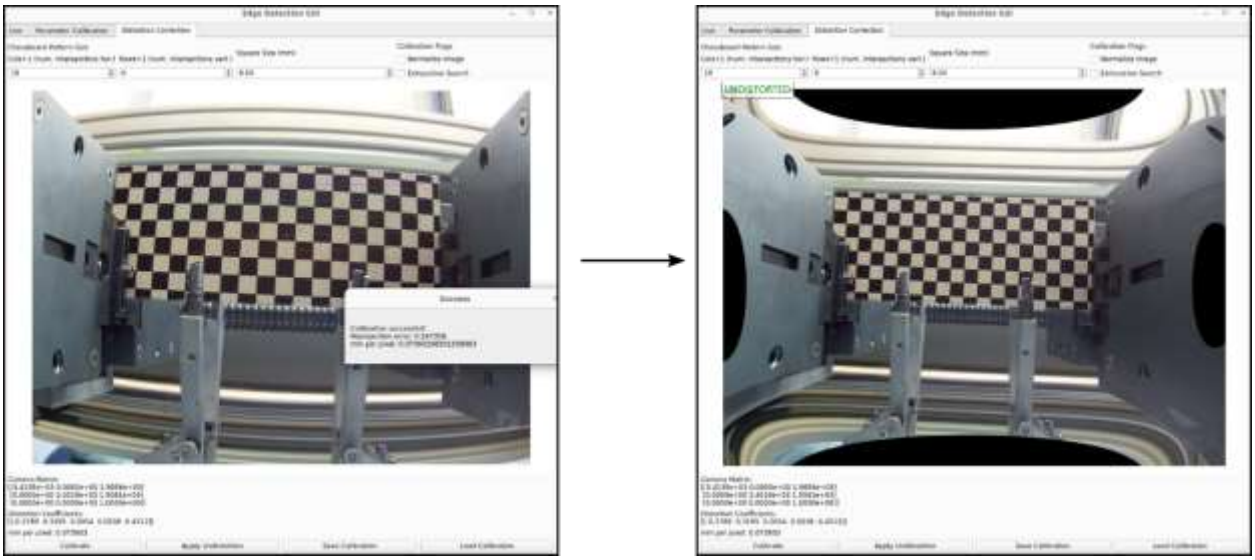

Abbildung 1. Kalibrierung der Kamera im GUI mittels Schachbrettmuster

Kantendetektion und Regelungsparameter

Überblick

Das entwickelte Verfahren besteht aus folgenden Bestandteilen: Canny-Edge-Detection mit anschließender Krümmungsfilterung, robuste Linienerkennung mittels RANSAC-Algorithmus, Abstandsberechnung zur Bestimmung der Probenbreite und abschließendes Mapping der Breitenveränderung auf Prozessparameter für Regelungsanwendungen. Diese Methoden werden sequenziell ausgeführt, um aus einem Eingangsbild Abstandsmessungen zu extrahieren und diese in steuerbare Ausgangssignale umzuwandeln.

Region of Interest Definition

Zur Reduktion der Rechenzeit und zur Fokussierung auf relevante Bildbereiche wird zunächst eine Region of Interest (ROI) definiert. Diese rechteckige Region wird durch zwei Eckpunkte $P_1(x_1,\ y_1)$ und $P_2(x_2,\ y_2)$ festgelegt und begrenzt den Arbeitsbereich des Algorithmus auf den Bereich, in dem die zu erkennenden Probenkanten erwartet werden.

Die ROI-Auswahl erfolgt basierend auf a-priori Wissen über die Anwendung und kann dynamisch durch den Nutzer im GUI angepasst werden. Dadurch werden nicht nur Rechenressourcen geschont, sondern auch Störeinflüsse aus anderen, irrelevanten Bildbereichen eliminiert. Ein Beispiel für eine so ausgewählte ROI ist in Abbildung 2 dargestellt.

Abbildung 2. Definition der Region of Interest (rotes Rechteck) im GUI. Lediglich dieser Bildbereich wird für die Kantendetektion ausgewertet

Canny-Algorithmus

Für die Kantendetektion wird der bewährte Canny-Edge-Detector eingesetzt, der sich unter anderem durch hohe Genauigkeit auszeichnet [3]. Der Algorithmus arbeitet mit zwei Schwellwerten: einem unteren und einem oberen Schwellwert, welche unter anderem bestimmen, wie stark der Helligkeitsgradient an einer Stelle sein muss, um als Kante klassifiziert zu werden.

Pixel mit Gradienten über dem oberen Schwellwert werden sofort als Kantenpixel klassifiziert, während Pixel zwischen den beiden Schwellwerten nur dann als Kanten betrachtet werden, wenn sie mit bereits identifizierten Kantenpixeln verbunden sind. Diese Hysterese-Schwellwertbildung führt zu geschlossenen Kantenzügen und reduziert fragmentierte Kanten.

Krümmungsbasierte Filterung

Um gerade Linien von gekrümmten Strukturen in naiver erster Instanz zu unterscheiden und auszufiltern, wird nach der Canny Kantendetektion eine Krümmungsfilterung durch Polynom-Fitting auf die erkannten Kanten angewendet. Dabei werden für jede Kontur die x- und y-Koordinaten extrahiert und ein Polynom dritten Grades folgender Form angepasst:

$$y = a_3x^3 + a_2x^2 + a_1x + a_0$$

Die Koeffizienten a_3 und a_2 charakterisieren die Krümmungseigenschaften der Linie. Der maximale Krümmungswert wird als Maximum der Absolutwerte dieser beiden Koeffizienten definiert:

$$\kappa = \max(|a_2|, |a_3|)$$

Konturen mit einem Krümmungswert unterhalb dieses definierten Maximalwerts, welcher ebenfalls im GUI konfigurierbar ist, werden als gerade Linien klassifiziert und für die weitere Verarbeitung beibehalten. Dieser Filterungsschritt stellt einen ersten naiven Schritt dar, um offensichtlich gekrümmte Strukturen zu eliminieren und fokussiert die Analyse auf die gewünschten geraden Kanten.

RANSAC-basierte Linienerkennung

Die Kanten der Zugproben werden als zwei linearen Geraden angenommen. Die robuste Erkennung dieser Linien beziehungsweise der Ermittlung der jeweiligen Geradengleichungen erfolgt mittels des RANSAC (Random Sample Consensus) Algorithmus, der besonders widerstandsfähig gegen Ausreißer ist [4]. Da in der Anwendung zwei horizontale, (parallele) Linien erwartet werden, wird die ROI horizontal in eine obere und untere Hälfte geteilt, und für jede Hälfte wird separat eine Gerade mittels RANSAC gefittet, welche die obere bzw. die untere Probenkante repräsentiert.

Der RANSAC-Algorithmus arbeitet iterativ: In jeder Iteration wird eine zufällige Teilmenge von mindestens zwei Punkten ausgewählt und eine Linie durch diese Punkte gelegt. Anschließend wird die Anzahl der Inlier-Punkte bestimmt, die innerhalb eines definierten Residualschwellwerts zur angepassten Linie liegen. Nach einer maximalen Anzahl von Versuchen wird das Modell mit den meisten Inliers als beste Lösung gewählt.

Jede Linie wird dann durch die Steigung k und den y-Achsenabschnitt d beschrieben. Für die untere Bildhälfte wird der y-Achsenabschnitt um die halbe Bildhöhe korrigiert, um die korrekte Position im Gesamtbild zu erhalten.

Abstandsberechnung

Die Abstandsmessung zwischen den beiden erkannten Linien erfolgt am horizontalen Mittelpunkt der ROI ($x^{mid} = w/2$, wobei w die Breite der ROI darstellt). An dieser Position werden die y-Koordinaten beider Linien berechnet:

$$y_1^{mid} = k_1 \cdot x^{mid} + d_1$$

$$y_2^{mid} = k_2 \cdot x^{mid} + d_2$$

Der Abstand zwischen den Linien ergibt sich als Absolutwert der Differenz dieser y-Koordinaten:

$$d = \left| y_2^{mid} - y_1^{mid} \right|$$

Abbildung 3 zeigt das Resultat einer so durchgeführten Kantenerkennung inklusive berechnetem Abstand (in Pixel), wobei rot markierte Linien durch RANSAC ausgeschlossene Kanten darstellen.

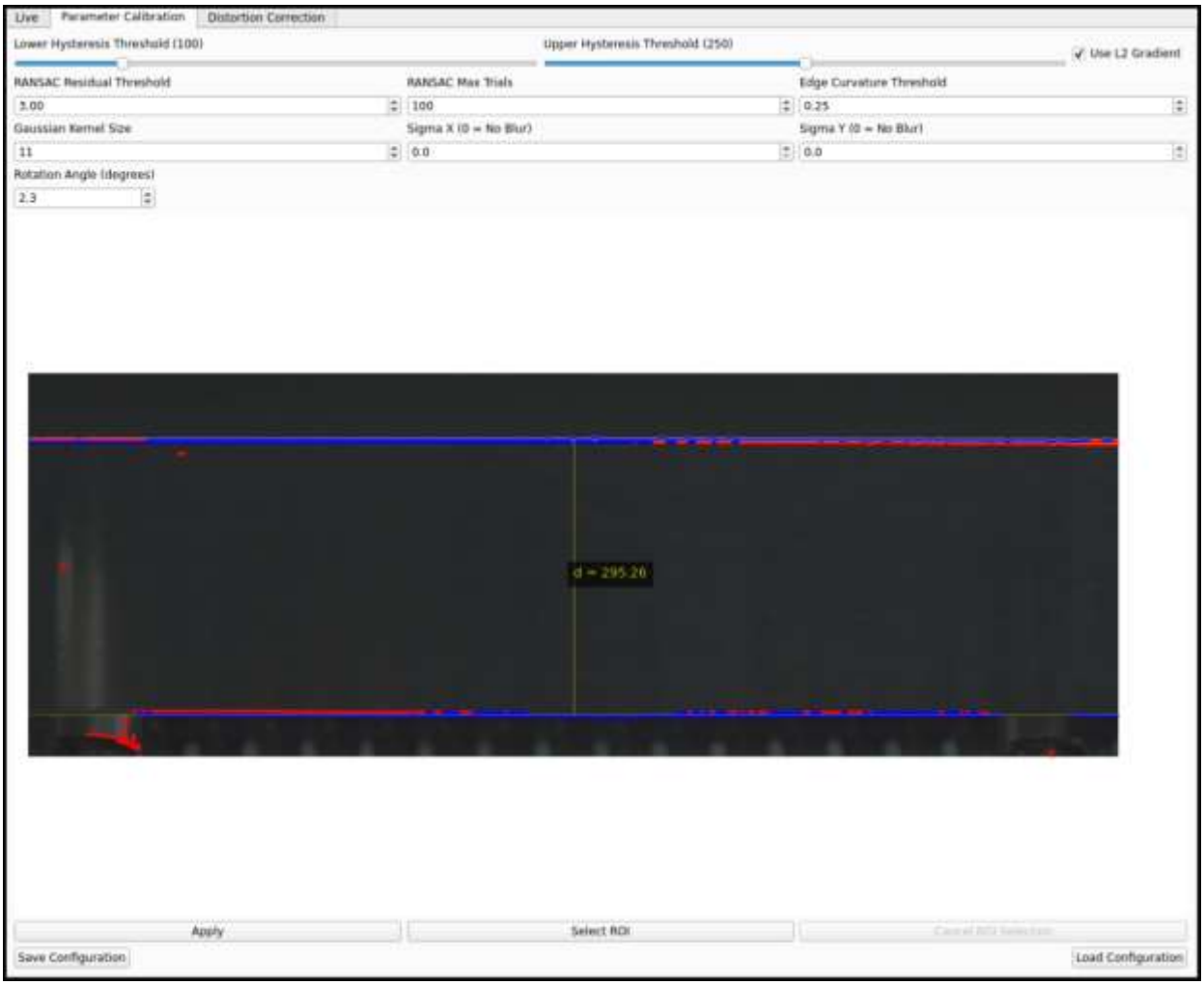

Abbildung 3. Ergebnis einer Kantendetektion. Die blauen Linien stellen die als Probenkanten klassifizierten Kanten dar

Mapping für Downstream Regelungsanwendung

Für die Integration in Regelungssysteme wird der gemessene Abstand d in ein Spannungssignal umgewandelt. Hierzu wird ein Referenzabstand d' (vor dem Ziehen der Probe) und eine entsprechende Referenzspannung U' definiert.

Die prozentuale Abweichung vom Referenzabstand wird dann berechnet als:

$$\Theta = \frac{(d - d')}{d'} \cdot 100$$

Mit Hilfe dieser Abweichung kann anschließend unter Verwendung eines Skalierungsfaktors α, der auch im GUI konfigurierbar ist, ein neuer Spannungswert vorgegeben werden, welcher dann der Regelungsanwendung übermittelt werden kann:

$$U_{new} = U' + \Theta \cdot \alpha$$

Abbildung 4 zeigt die Live-Ansicht des GUI, welche während des Ziehens der Probe angezeigt wird. Darin werden sowohl die erkannten Probenkanten als auch die errechnete Veränderung des Kantenabstands zum Referenzabstand vor dem Ziehen angezeigt. Zudem kann direkt im GUI ein Faktor

für das Spannungs-Mapping eingestellt werden, um die Distanzänderung linear in eine auszugebende Regelgröße zu verwandeln.

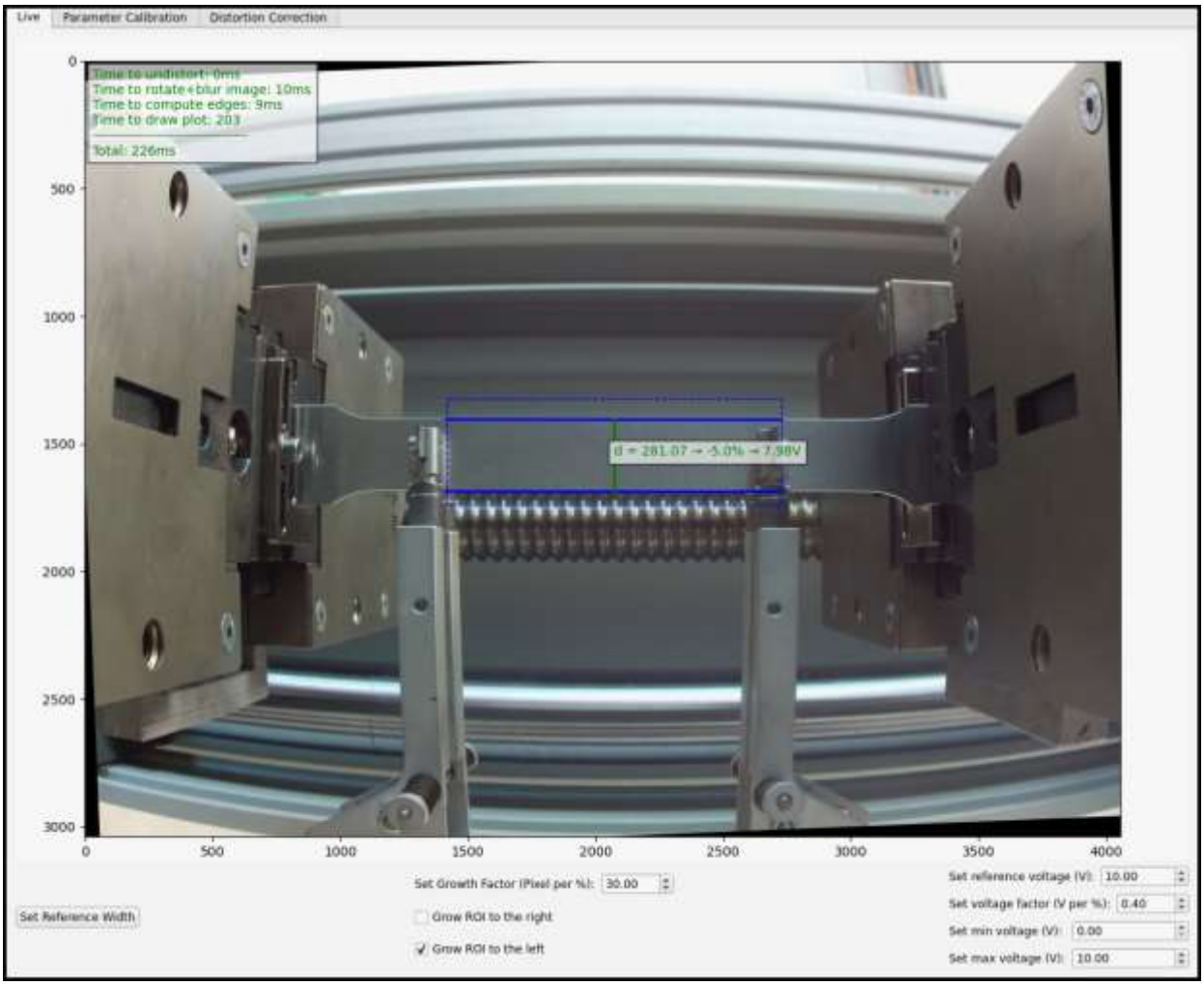

Abbildung 4. Zeigt die Live-Ansicht im GUI während des Ziehens einer Zugprobe, inklusive erkannten Kanten in blauen durchgezogenen Linien sowie Linienabstand in Pixel und prozentueller Änderung im Vergleich zum Referenzwert vor dem Ziehen und daraus resultierender Regelgröße (Spannung)

Ergebnis

Das System wurde erfolgreich an der semi-industriellen Anlage des LKR validiert. Die plastische Verformung der Probe konnte durch das optische Modul in Echtzeit identifiziert und quantifiziert werden. Auf Basis der erfassten Verdünnungswerte wurde der Stromregelpfad dynamisch angepasst, indem gemappte Steuerungsparameter in die Ansteuerung der Hochstromquelle eingespeist wurden. Dadurch könnte in weiterer Folge die Rissbildung gezielt verzögert und die Umformbarkeit verbessert werden.

Die Kombination aus robuster, bildgestützter Geometriemessung und geringer Systemlatenz macht die Lösung besonders attraktiv für kleine und mittlere Unternehmen mit variablen Umformprozessen und begrenztem Sensoreinsatz.

Die Innovation der Lösung liegt in der Umsetzung einer echtzeitfähigen, bildbasierten Closed-Loop-Regelung des elektroplastischen Effekts (EPE) in metallischen Umformprozessen. Durch die optische Erfassung der plastischen Verdünnung per Canny- und RANSAC-Analyse wird die Stromimpulssteuerung dynamisch angepasst, sodass die EPE-induzierte Umformbarkeit gezielt beeinflusst werden kann. Die Lösung arbeitet latenzarm auf Edge-Hardware, ist modular skalierbar und

bildet die Grundlage für zukünftige KI-gestützte Regelungen in mehrachsigen Tiefziehprozessen mittels digitaler Bildkorrelation (DIC).

In einem geplanten Nachfolgeprojekt soll das System für komplexe, mehrachsige Tiefziehprozesse erweitert werden. Dabei ist der Einsatz digitaler Bildkorrelation (DIC) zur mehrachsigen Dehnungserfassung vorgesehen, ergänzt um ein generisches, lernfähiges Regelmodul auf Edge-Hardware. Ziel ist die Entwicklung eines selbstoptimierenden, adaptiven Steuerungssystems für den intelligenten Betrieb zukünftiger Umformanlagen.

Zusammenfassung

Der elektroplastische Effekt (EPE) ermöglicht die Reduktion von Umformkräften und die Erhöhung der Duktilität in metallischen Werkstoffen. Bisherige Verfahren arbeiteten jedoch mit starren Stromprofilen ohne Rückkopplung auf die tatsächliche plastische Verformung. Der vorgestellte Beitrag entwickelt die Basis für ein echtzeitfähiges, bildbasiertes Closed-Loop-Regelungssystem, das über optische Erfassung der Probengeometrie eine adaptive Steuerung der Stromimpulse ermöglicht.

Das System integriert Kamerakalibrierung, Canny-Edge-Detection, RANSAC-basierte Linienerkennung und Echtzeitvisualisierung in einer GUI-Anwendung. Über die kontinuierliche Überwachung der Probenverdünnung, welche in diesem Beitrag entwickelt wurde, können später Steuerungsparameter dynamisch angepasst und in die Hochstromquelle eingespeist werden. Die Kombination aus robuster Bildanalyse und geringer Systemlatenz ermöglicht präzise Regelungseingriffe während des Umformprozesses. Die erfolgreiche Validierung an der semi-industriellen Anlage des LKR bestätigte die Praxistauglichkeit des Systems, wobei die Verformung in Echtzeit identifiziert und quantifiziert werden konnte.

Die entwickelte Lösung bietet auch Übertragungsmöglichkeiten auf andere Industriezweige: Das Prinzip der bildbasierten Closed-Loop-Regelung ist auf verschiedene Umformverfahren übertragbar, wobei kleine und mittlere Unternehmen besonders von der kosteneffizienten Edge-Hardware-Lösung ohne aufwendige Sensortechnik profitieren. Zukünftige KI-gestützte Erweiterungen mit digitaler Bildkorrelation eröffnen Potenziale für selbstoptimierende Produktionssysteme in verschiedensten Fertigungsbereichen.

Literaturverzeichnis

[1] Computing, Riverbank, „PyQt6,“ [Online]. Available: https://www.riverbankcomputing.com/software/pyqt/. [Zugriff am 28 07 2025].

[2] „OpenCV - Github,“ [Online]. Available: https://github.com/opencv/opencv-python. [Zugriff am 28 07 2025].

[3] J. Canny, „A Computational Approach to Edge Detection,“ *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Bde. %1 von %2PAMI-8, Nr. 6, pp. 679-698, 1986.

[4] R. C. B. Martin A. Fischler, „Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography,“ *Communications of the ACM,* Bd. 24, Nr. 6, pp. 381-395, 1981.

Digitale mensch-zentrierte Assistenzsysteme

Christoph Heindl
Visual Computing
PROFACTOR GmbH
cheind@profactor.at

Gerhard Ebenhofer
Visual Computing
PROFACTOR GmbH
gebenh@profactor.at

Gernot Stübl
Visual Computing
PROFACTOR GmbH
gstueb@profactor.at

Harald Bauer
Visual Computing
PROFACTOR GmbH
hbauer@profactor.at

Kurzzusammenfassung

Der zunehmende Fachkräftemangel, steigende Variantenvielfalt und der Wunsch nach ergonomischen Arbeitsplätzen erfordern neue Ansätze für mensch-zentrierte Automatisierungslösungen. In diesem Beitrag wird eine modulare Toolbox vorgestellt, welche KI-basierte Wahrnehmung, projektive Interaktion und Robotik kombiniert, um adaptive Assistenzsysteme in variantenreichen Produktionsumgebungen zu realisieren. Zentrale Bestandteile sind die kamerabasierte Erkennung von Objekten und Personen sowie orts- und zeitgenaue Projektionen zur direkten Arbeitsplatzunterstützung. Für die Auftragserstellung stehen intuitive Modalitäten zur Verfügung: wie die Stifteingabe oder SketchGuide+, ein freihändiges Skizziersystem mit Wearable und visueller Rückmeldung. Die praktische Machbarkeit wurde in mehreren TBI-Projekten in Branchen wie Möbelbau, Sanitär- und Metalltechnik oder Solarproduktion demonstriert. Die Ergebnisse zeigen hohes Potenzial für Prozessverbesserung, Benutzerfreundlichkeit und Flexibilität in manuellen oder teilautomatisierten Abläufen.

Keywords: Digitale Assistenz, Visuelles Szenenverständnis, Projektion, Robotik

Abstract

The increasing shortage of skilled workers, growing variety of products and the desire for ergonomic workplaces require new approaches to human-centred automation solutions. In this work, we present a modular toolbox that combines AI-based perception, projective interaction and robotics to implement adaptive assistance systems in diverse production environments. Key components include camera-based recognition of objects and people, as well as location- and time-specific projections for direct workplace support. Intuitive modalities are available for order creation, such as pen input or SketchGuide+, a freehand sketching system with wearable technology and visual feedback. The practical feasibility has been demonstrated in several TBI projects in industries such as furniture manufacturing, sanitary and metal technology, and solar production. The results show significant potential for process improvement, user-friendliness, and flexibility in manual or semi-automated processes.

Keywords: Digital Assistance, Visual Scenes, Projection, Robotics

Einleitung

Die zunehmende Komplexität industrieller Prozesse, der steigende Bedarf an Produktindividualisierung sowie der fortschreitende Fachkräftemangel stellen produzierende Unternehmen vor wachsende Herausforderungen. Mensch-zentrierte Assistenzsysteme bieten hier vielversprechende Ansätze, um Arbeitsprozesse flexibler, ergonomischer und effizienter zu gestalten, ohne den Menschen aus dem Zentrum der Tätigkeit zu verdrängen.

Bestehende Assistenzsysteme fokussieren sich häufig auf stark strukturierte Szenarien, in denen z.B. Projektionsmuster nach starren Abläufen gezeigt werden und keinerlei Anpassung an die Arbeitsweise von Werker:innen geschieht. Oft setzen sie auch eine vollständige Automatisierung voraus, welche bei Klein- und Mittelbetrieben selten zu finden ist. Zudem kommt noch, dass meist die Bildschirme oder Touchpanels vom realen Arbeitskontext räumlich und funktional getrennt sind, was eine hohe kognitive Belastung von Werker:innen durch den Medienbruch verursacht. Zwar existieren bereits erste Systeme mit Augmented Reality-Displays oder sensorgestützter Robotik, doch deren Konfiguration und Programmierung erfordert vielfach spezialisiertes Fachwissen. Die Integration in variantenreiche, manuell geprägte Produktionsumgebungen bleibt oft begrenzt.

Der vorgestellte Ansatz verfolgt einen anderen Weg: die entwickelte modulare Toolbox, welche KI-basierte Wahrnehmung, projektive Interaktionstechnologien und adaptive Robotik kombiniert, bringt Assistenz dort, wo sie unmittelbar benötigt wird, siehe Abbildung 1. Also direkt auf dem Werkstück, in der Umgebung und im realen Arbeitsfluss. Zentrale Elemente sind die kamerabasierte Erkennung von Objekten, Personen und Kontextinformationen sowie die orts- und zeitgenaue Projektion von Handlungsempfehlungen oder Eingabeflächen.

Abbildung 1. Vielfältige Einsatzszenarien der vorgestellten Toolbox für mensch-zentrierte Assistenzsysteme: von projektion-basierter Montageunterstützung und robotergestütztem Polieren sowie Qualitätssicherung bis hin zu KI-gestützter Objekterkennung und freihändiger Auftragsdefinition mit SketchGuide+.

Für die Auftragseingabe stehen zwei gleichwertige, intuitive Modalitäten zur Verfügung: die klassische Stifteingabe mit projektiver Rückmeldung sowie SketchGuide+ [1,2], ein System zur freihändigen Skizzierung auf realen Oberflächen, erweitert um ein tragbares Wearable zur Echtzeit-Parametereingabe und zeitlich synchronisierte Projektionen. Beide Methoden ermöglichen eine niederschwellige Interaktion, auch für nicht-technische Nutzer:innen, und erlauben eine schnelle Anpassung an wechselnde Anforderungen.

Die technische und ergonomische Machbarkeit dieses Meta-Ansatzes wurde in mehreren branchenspezifischen Kurzprojekten validiert. In Bereichen wie Möbelbau, Sanitärproduktion, Solarrahmenfertigung oder Logistik konnten Potenziale für verbesserte Prozessqualität, reduzierte Rüstzeiten und gesteigerte Nutzerakzeptanz demonstriert werden.

Vorgehensweise mit Fokus auf risikobehaftete Kernfragen

Das methodische Vorgehen in den erfolgreichen TBI-Projekten folgte einem strukturierten, praxisnahen Schema, das darauf abzielte, die größten technologischen Unsicherheiten frühzeitig und unter realistischen Bedingungen zu untersuchen. Im Zentrum stand dabei nicht die vollständige Umsetzung eines Gesamtsystems, sondern die gezielte Validierung risikobehafteter Kernfragen mit hoher Relevanz für eine spätere Implementierung. Das Vorgehen umfasste:

1. Frühe Datenerfassung im Realbetrieb:
   Bereits zu Projektbeginn wurden reale Teiledaten und Prozessinformationen direkt vor Ort aufgenommen, um ein belastbares Verständnis des tatsächlichen Anwendungskontexts zu erlangen.

2. Fokussierung auf kritische Machbarkeitsaspekte:
   Statt breiter Anforderungsanalysen wurden die jeweiligen Problemstellungen konsequent auf wenige, aber zentrale Unsicherheiten reduziert, etwa die Qualität eines robotischen Polierprozesses, die Robustheit eines KI-Modells oder die intuitive Bedienbarkeit einer Interaktionsschnittstelle.
3. Gezielte Auswahl und Adaption der Toolbox-Komponenten:
   Je nach Risikoprofil wurden passende Module aus der Toolbox ausgewählt (z. B. KI-gestützte Bildverarbeitung, projektionsbasierte Assistenz, Interaktion über Stift oder SketchGuide+) und projektspezifisch angepasst, etwa durch Finetuning künstlicher neuronaler Netze oder Hardwareintegration.
4. Schlanke Systemintegration:
   Um den Aufwand zu begrenzen, erfolgte die Anbindung an betriebliche Informationssysteme in möglichst einfacher Form, ausreichend für realitätsnahe Tests, ohne komplexe IT-Projekte auszulösen.
5. Demonstratoraufbau und Bewertung:
   Die resultierenden Prototypen wurden entweder im Labor oder direkt beim Partnerunternehmen installiert und unter realen Bedingungen qualitativ wie quantitativ bewertet, mit Fokus auf Zielgrößen wie Prozessqualität, Interaktionszeit, Fehlerrobustheit oder Nutzerakzeptanz.

Dieses risikoorientierte Vorgehen ermöglichte es, zentrale Implementierungshürden frühzeitig zu adressieren, belastbare Aussagen zu technologischer Machbarkeit zu treffen und gleichzeitig eine fundierte Grundlage für Folgeprojekte oder Konzepterstellungen zu schaffen.

Ergebnisse

Im Gegensatz zu konventionellen Assistenzsystemen, die oft auf isolierte Interaktionsmethoden, starre Automatisierungslogiken oder rein bildschirmbasierte Rückmeldungen setzen, verfolgt der vorgestellte Ansatz einen ganzheitlich integrierten Weg: Die Toolbox kombiniert KI-basierte Umgebungswahrnehmung, adaptive Robotik, und projektiongestützte, kontextsensitive Mensch-Maschine-Interaktion zu einem flexiblen Baukasten für mensch-zentrierte Assistenzsysteme.

Die zentralen Innovationsmerkmale sind:

1. Direkte Interaktion im realen Arbeitskontext:
   Statt virtuellem Interface werden Informationen dort eingeblendet, wo sie gebraucht werden, also direkt auf Objekten, Oberflächen oder Werkstücken.
2. Multimodale Auftragseingabe durch Stift oder SketchGuide+:
   Nutzer:innen können Roboterpfade entweder durch klassische Stifteingabe oder mittels SketchGuide+ freihändig auf der realen Oberfläche definieren, inklusive tragbarer Parametereingabe und sofortiger visueller Rückmeldung.
3. KI-gestützte Echtzeitwahrnehmung für variantenreiche Objekte und Szenarien:
   Objektformen, Pakete oder Auftragslayouts werden automatisiert erkannt, analysiert und für robotische Aktionen aufbereitet, ohne explizite Programmierung.
4. Einfacher Technologietransfer über TBI
   Die Toolbox wurde systematisch in unterschiedlichen Domänen wie Sanitärtechnik, Möbelbau, Logistik und Solarproduktion getestet, wodurch sowohl technologische als auch ergonomische Machbarkeit nachgewiesen wurde.

Durch diese modulare Kombination entsteht ein leistungsfähiges Assistenzsystem, das auch bei Einzelstückfertigung, häufigem Variantenwechsel und für nicht-technischen Anwender:innen eingesetzt werden kann, mit hoher Adaptivität, minimalem Rüstaufwand und intuitiver Bedienbarkeit.

Insbesondere konnten unsere Technologien in folgenden Aufgabenstellungen evaluiert werden:

- Ein teilautomatisierter Polierprozess für Waschbecken wurde mithilfe von Robotik, 3D-Scan und intuitiver Auftragseingabe (Stift oder SketchGuide+) umgesetzt, um Ergonomie und Qualität bei Einzelstückbearbeitung zu verbessern.
- Für den situativen Kleberauftrag im Möbelbau wurde ein robotergestütztes System mit Bildverarbeitung und benutzerfreundlicher Eingabe (Touch oder SketchGuide+) realisiert, das individuelle Aufträge in Sekunden ermöglicht.
- Ein AR-basiertes Entladesystem projiziert visuelle Anweisungen direkt auf Blechteile, reduziert Papierprozesse und verbessert die Effizienz beim Sortieren und Entnehmen.
- Der Schweißprozess individueller Solarkollektorrahmen wurde durch Roboterführung und Pfadeingabe via Stift oder SketchGuide+ teilautomatisiert, wodurch die Rüstzeit für Einzelanfertigungen deutlich sank.
- Durch digitale Projektionen konnten physische Auflegeschablonen beim Bootsbau ersetzt und damit Auflegezeiten sowie Lageraufwand signifikant reduziert werden.
- Ein KI-basiertes System zur Paketklassifikation und -verfolgung demonstrierte, dass visuelle Erkennung auch unter schwierigen Lichtverhältnissen und in Echtzeit zuverlässig funktioniert.
- Die manuelle Ausfertigung von Mineralguss-Waschtischen wurde erfolgreich durch ein kollaboratives Robotersystem mit nutzerdefinierter Bearbeitungslogik unterstützt und in einer realen Produktionsumgebung validiert.

Tabelle 1 gruppiert die Technologien nach Komponenten.

Tabelle 1. Toolbox: Technologien im praktischen Einsatz.

Toolbox-Komponente	Beispielhafte Anwendungen
Objekt- & Menschenerkennung mit Kamera & KI	Bildverarbeitung & Segmentierung für individuelle Teile (Kleberauftrag, Paketannahme, Polieren)
Projektionstechnologie zur AR-Interaktion	Digitale Auflegeschablonen, visuelle Entladeanweisungen, projizierte Schweißpfade
Roboter-Assistenzsysteme	Teilautomatisierte Politur, Schweißen, Finishing, Kleberauftrag
Multimodale Interaktion (Touch / Stift / Handführung)	Bedienung via Stifteingabe, touchbasierte Interfaces, gestenbasierte Programmierung
Digitale Assistenz & visuelle Feedbacksysteme	Projektion, API-Schnittstellen, Integration in reale Prozesse

Abbildungen 2 bis 4 zeigen ausgewählte Bilder aus spezifischen TBIs.

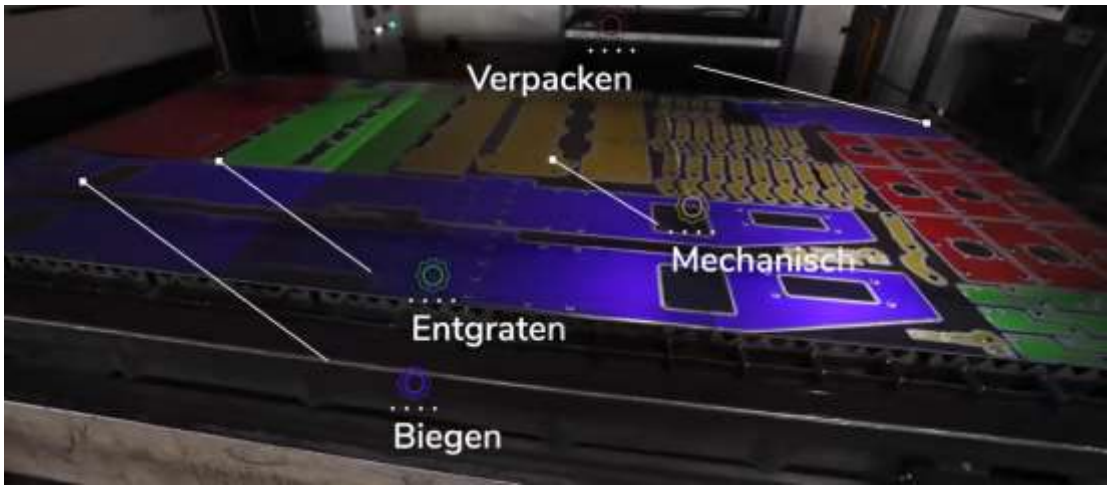

Abbildung 2. Digitaler Entladeassistent im Einsatz: Projektion von visuellen Entladehinweisen direkt auf Blechteile zur Unterstützung manueller Sortierprozesse, ermöglicht durch ein Multi-Projektor-System und leichtgewichtiger Integration in betriebliche Abläufe.

Abbildung 3. Digitaler Assistent für die Paketannahme: KI-basierte Objekterkennung identifiziert und verfolgt Pakete zuverlässig auch bei schlechten Lichtverhältnissen. Eine schlanke Lösung für offene Regalsysteme ohne manuelle Zuordnung.

Abbildung 4. Digitale Auflegeschablonen im Bootsbau: Projektion ersetzt physische Schablonen und ermöglicht ein präzises, zeiteffizientes Auflegen von Stabdecks direkt auf der Werkfläche.

Zusammenfassung

Es wurde eine Toolbox für mensch-zentrierte Assistenzsysteme vorgestellt, welche KI-basierte Wahrnehmung, Robotik und projektiongestützte Interaktion kombiniert. In mehreren TBI-Projekten wurde sie praxisnah eingesetzt, um risikobehaftete Kernfragen frühzeitig zu validieren. Durch reale Datenerfassung, gezielte Anpassung der Komponenten und schlanke Prototypen konnten technologische Machbarkeiten in verschiedenen Domänen demonstriert werden. Besonders die Kombination aus kontextsensitiver Projektion und intuitiver Eingabe, per Stift oder SketchGuide+, erwies sich als Schlüssel zur flexiblen Unterstützung manueller Prozesse und zur fundierten Bewertung von Umsetzungspotenzialen.

Die dargestellte Technologie ist im Gegensatz zu anderen Lösungen nicht auf den Einsatz in Großbetrieben limitiert, sondern bringt auch Klein- und Mittelbetrieben direkte Zeitersparnis und erhöhte Qualitätssicherung durch Unterstützung bei manuellen Tätigkeiten. Gerade die Flexibilität der Toolbox ermöglicht es, sich entweder auf geänderte Fertigungsanforderungen oder auch an die Arbeitsweise der Werker:innen anzupassen. Dies ist ein deutlicher Vorteil gegenüber anderen Systemen. Die Technologie ist branchenunabhängig konzipiert und richtet sich an die noch stark vorherrschenden manuellen Arbeitsplätze in österreichischen Produktionsumgebungen.

Literatur

[1] Heindl, Christoph, et al. "SketchGuide: A Baseline Vision-Based Model for Rapid Robot Programming via Freehand Sketching on Any Surface." European Robotics Forum. Cham: Springer Nature Switzerland, 2025.

[2] Heindl, Christoph, et al. "SketchGuide+: Enabling Parametric Freehand Robot Programming via Wearable-Assisted Sketching." ACM International Joint Conference on Pervasive and Ubiquitous Computing, Espoo, Finland, 2025.

Intelligentes Assistenzsystem für flexible Mensch-Roboter-Kollaboration in der Fertigung

Helmut Nöhmayer
PROFACTOR GmbH
helmut.noehmayer@profactor.at

Christian Wögerer
PROFACTOR GmbH
christian.woegerer@profactor.at

Oswald Bratu
PROFACTOR GmbH
oswald.bratu@profactor.at

Andreas Pichler
PROFACTOR GmbH
andreas.pichler@profactor.at

Kurzzusammenfassung

Der steigende Bedarf an individualisierten Produkten, ergonomischen Arbeitsplätzen sowie der zunehmende Fachkräftemangel erfordern neue humanzentrierte Automatisierungsansätze. Dies gilt besonders für Anwendungen mit hoher Variantenvielfalt und geringer Stückzahl. In diesem Beitrag wird ein intelligentes Assistenzsystem vorgestellt, das KI-unterstütze Wahrnehmung, adaptive Robotik und multimodale Interaktionstechnologien kombiniert. Ziel ist es dabei, die Mensch-Roboter-Kollaboration (MRK) auch in manuell geprägten Fertigungsprozessen praktikabel und wirtschaftlich zu gestalten. Das System ermöglicht die intuitive Auftragserstellung mittels Stifteingabe, Gestik oder Sprache direkt am Werkstück. Durch die situative, automatische Generierung und Optimierung von Roboterbahnen können selbst Einzelstücke effizient bearbeitet werden – ohne klassischen Programmieraufwand. Die technische Machbarkeit wurde im Rahmen mehrerer industriebezogener Anwendungsfälle validiert. Die Ergebnisse zeigen signifikante Potenziale für die Reduktion von Rüstzeiten und bei Steigerung der Produktqualität, Ergonomie und Arbeitsplatzattraktivität

Keywords: Mensch-Roboter-Kollaboration, Assistenzsystem, KI-basierte Prozessplanung, Stifteingabe, flexible Automation

Abstract

The increasing demand for customized products, ergonomic workplaces, and the growing shortage of skilled labor call for new human-centered automation approaches. This is particularly relevant for applications characterized by high product variability and low production volumes. This paper presents an intelligent assistance system that integrates AI-based perception, adaptive robotics, and multimodal interaction technologies. The objective is to make human-robot collaboration (HRC) both practical and economically viable, even in manufacturing processes that are predominantly manual.

The system enables intuitive task specification directly at the workpiece through pen input, gestures, or speech. By automatically generating and optimizing robot trajectories in situ, even single-piece production can be carried out efficiently—without the need for traditional programming. The technical feasibility has been validated through several industry-related application scenarios. The results demonstrate significant potential for reducing setup times while improving product quality, ergonomics, and workplace attractiveness.

Keywords: Human-Robot Collaboration, Assistance System, AI-Based Process Planning, Pen Input, Flexible Automation

Einleitung

In der industriellen Fertigung wächst der Druck, sowohl auf individuelle Kundenanforderungen flexibel zu reagieren und auch gleichzeitig Qualität, Effizienz und Ergonomie sicherzustellen und zu steigern. Prozesse wie Polieren, Kleben oder Schweißen sind dabei oft stark durch monotone, manuelle

Bearbeitung geprägt, insbesondere bei Einzel- oder Kleinserienfertigung. Diese Tätigkeiten sind nicht nur zeit- und personalintensiv, sondern bergen auch gesundheitliche Risiken und sind zunehmend schwerer zu besetzen.

Klassische Automatisierungssysteme bieten hierfür nur begrenzt Lösungen und die notwendige betriebliche Flexibilität: Sie folgen teilweise starren Programmierlogiken und bedingen damit lange Rüstzeiten im Verhältnis zu den Prozesszeiten. Darüber hinaus benötigen sie spezielle Fachkenntnisse In der Praxis führt dies dazu, dass viele manuelle Prozesse bislang nicht automatisiert werden– trotz offensichtlicher Potenziale zur Effizienzsteigerung, da die Programmierzeiten den Zeitbedarf für die manuelle Ausführung teilweise erheblich überschreitet.

Inhalt dieses Beitrags ist die Vorstellung eines intelligenten, modularen Assistenzsystems, das durch KI-basierte Wahrnehmung, sensorgestützte Aufspannungserkennung und intuitive Mensch-Maschine-Interaktion eine neue Qualität der Mensch-Roboter-Kollaboration ermöglicht. Dabei steht nicht eine vollständige Automatisierung im Vordergrund, sondern die gezielte Teilautomation monotoner Prozessschritte unter Beibehaltung der menschliche Steuerungshoheit und Erfahrung in Bezug auf Ausführungsqualität und Ergebnis.

Abbildung 1. Auftragserzeugung für Spotbearbeitung von Waschtischen

Technologische Zielstellung

Das System zielt auf eine radikale Reduktion von Auftragserzeugungszeiten bei der Variantenbildung ab. Durch den Einsatz einer multimodalen Bedieninterfaces und situativer Visualisierung des Auftrags und des Ausführungsfortschritts (Stift, Gestik, Sprache, Touch, Projektion) können Bearbeitungsaufträge direkt am realen Werkstück situationsbezogen sehr rasch erzeugt werden. Ermöglicht wird dies durch:

- Lage- und Geometrieerkennung mittels 2D +3D Kamerasystemen mittel zusammengesetzter Bilder von großflächigen/großvolumigen Arbeitsräumen
- KI-gestützte Erfassung der individuellen Teileaufspannung / Teilebereitstellungen
- Intuitiver und rasche Eingabe von Einsatzpunkten und Bearbeitungsregion über Handgesten und getrackte Zeigeinstrumente für eine individuelle Spotbearbeitung
- Automatisierte Bahnplanung und Prozesserzeugung für diese situativen Bearbeitungsaufträge
- Projektionsunterstützte Visualisierung und Rückmeldung zur Validierung der Bearbeitungspfade direkt am Produkt

Die Kombination dieser Methoden und Komponenten ermöglichen eine Auftragserstellung innerhalb weniger Sekunden pro Bearbeitungsstelle (Spot) – auch für Unikatanwendungen und Losgröße 1 Produkte, siehe Abbildung 1. Die eigentliche Auftragserzeugung erfolgt vollständig automatisiert im Hintergrund. Dadurch wird der Einsatz von Robotik erstmals auch in Umgebungen mit hoher Variantenvielfalt und geringer Automatisierungskompetenz praktikabel. Wirtschaftlich unterstützt dieser Ansatz die Reduktion von Rüstaufwänden / Stillstandszeiten von Maschinen und dient damit der Senkung der Lohnstückkosten im Hinblick auf die Wettbewerbsfähigkeit, siehe Abbildung 2.

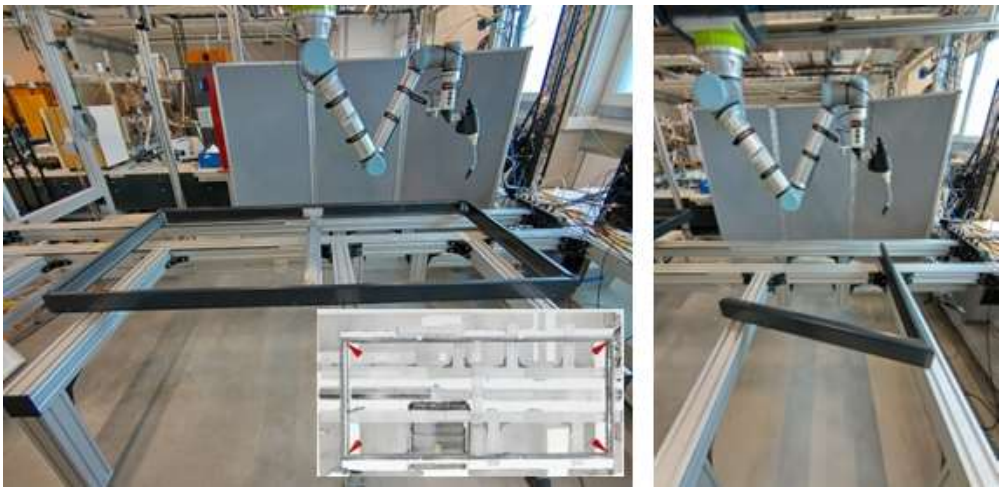

Abbildung 2. 3D Erfassung von Kollektorrahmen

Vorgehensweise: Validierung risikobehafteter Kernaspekte

Die Entwicklung und Validierung des Systems folgte konkret identifizierter industrieller Herausforderungen in mehreren TBI-Projekten. Das methodische Vorgehen war auf die Analyse und Evaluierung der technischen Risiken und die Ableitung und Bewertung der konzeptionellen Grenzen ausgerichtet und fokussierte sich hierfür auf folgende zentrale Fragestellungen:

- Polierprozesse:
  - Ist es möglich, mit einem Roboter gleichbleibende Hochglanzqualität bei variantenreichen Waschbeckenformen?
  - Ist eine automatische Erfassung der Beladesituation in 3D so möglich, dass daraus Bearbeitungsregionen für die Programmierung / Auftragserzeugung auch bei verschiedene Beckenformen identifiziert werden können?
- Schweißprozesse:
  - Kann die Geometrie individueller Solarrahmen automatisiert erkannt und daraus in Echtzeit ein robuster individueller Schweißauftrag abgeleitet werden?
  - Welche Abfolge in der Lageerfassung und Positionskorrekturen sind für einen Ad-Hoc - Schweißprozesseingabe für Individualbauteile wie z.B. die angewinkelten Aufständerungen von Solarrahmen für die Montage an den Dächern notwendig?
- Auftragseingabe:
  - Wie lässt sich eine intuitive Auftragserzeugung für ein Unikatprodukt / einen Unikatprozess ohne Programmierkenntnisse so realisieren, dass sie in einem minimalen Zeitfenster erfolgen kann und besonders einfach erlernbar ist?
- Zur Beantwortung dieser Fragen wurden realitätsnahe Funktionsmuster aufgebaut, die unter anderem mit kollaborativen Robotern, instrumentierten Werkzeugen, 3D-Kamerasystemen sowie projektionsbasierten Feedback-Komponenten ausgerüstet waren, siehe Abbildung 3.
- Die Evaluation dieser Systeme erfolgte sowohl im Labor als auch in realen Produktionsumgebungen.

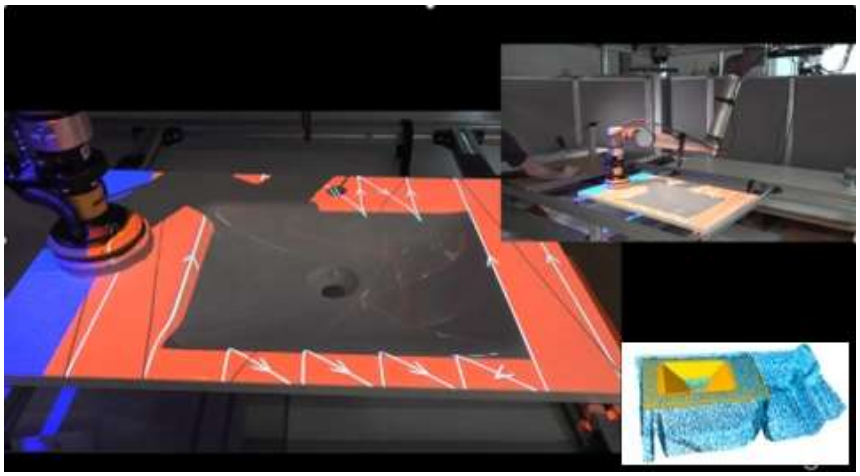

Abbildung 3. 3D Erfassung von Waschtischen

Ergebnis

Die Ergebnisse der Validierung zeigen, dass das entwickelte System in der Lage ist, zentrale Anforderungen an eine intuitive Nutzung von Robotersystemen für Unikatprozesse zu erfüllen:

- Rüstzeitersparnis: Auftragseingabenzeiten bewegen sich je nach Anwendung und Prozess im Bereich von 10-30s / Prozesspunkt
- Qualität: Die Ausführung der Bearbeitung durch eine Robotersystem liefert konstantere Ergebnisse auch bei variierenden Geometrien und Aufspannungen.
- Benutzerfreundlichkeit: Die intuitive Interaktion via Touch, Stift, Gestik und Projektion wurde von Anwender:innen als leicht erlernbar und anwendbar bewertet
- Skalierbarkeit: Die Lösung kann auf unterschiedliche Fertigungsprozesse (Polieren, Schleifen, Schweißen, Kleben, Ölen und auch Lackieren) ohne grundlegende Anpassungen übertragen werden.

Die technische Machbarkeit wurde somit unter realen Bedingungen erfolgreich demonstriert. Besonders hervorzuheben ist die Kombination aus geometrischer Echtzeiterkennung und intuitiver Auftragserstellung, die eine Automatisierung auch bei Einzelstücken wirtschaftlich ermöglicht.

Innovation und Systemarchitektur

Der Innovationsgrad des Assistenzsystems liegt insbesondere in der durchgängigen Integration folgender Teiltechnologien:

- 2D/3D Lagerfassung von Beladungssituationen aus zusammengesetzten Bildern großer Arbeitsräume [1].
- Instrumentierte Werkzeuge und Sensorik zur Eingabe von Position, Lage und Bearbeitungsintention.
- Multimodale Interaktion: Kombination von Touchscreen, Stifteingabe [2], Sprachsteuerung, Gestenerkennung und Führung / Feedback über Projektion.
- KI-unterstützte Prozessplanung: Automatische Umwandlung von Nutzerinteraktionen in robotergestützten, situationsabhängigen Aktionspfaden [3][4].
- Lernfähige Parametrierung: Automatische Anpassung an Material, Geometrie und Prozesszustand.

Im Zusammenspiel dieser Komponenten entsteht ein System, das klassische Programmierlogik durch situatives Handeln ersetzt – und damit einen Paradigmenwechsel in der industriellen Robotik einleitet. Dieser Lösungsansatz adressiert dabei im Besondern die Bedürfnisse von KMU in Bezug auf die Erschließung von Automationspotentialen für eine weiterhin wirtschaftliche Fertigung in Europa.

Nutzen und Implikationen für die Fertigung

Die entwickelte Lösung bietet ein breites Spektrum an Vorteilen für produzierende Unternehmen, insbesondere im Kontext von Fachkräftemangel, Losgröße-1-Fertigung und digitaler Transformation:

- Breite Anwendbarkeit bei typischen manuelle Bearbeitungsvorgänge in der Fertigung in den Bereichen der Holz-, Metall- und Kunststoffverarbeitung wie Schleifen, Bürsten, Polieren, Ölen, Lackieren Schweißen oder Kleben.
- Interaktive, rasche Auftragserzeugen für Robotersystem für Produktevarianten und Prozessvariationen die bisher ausschließlich durch den Menschen und seine Expertise bestimmt wurden - z.B. Rasche Eingabe eines Schleifauftrags bei Zwischenschliffen nach Begutachtung durch die Expert:innen und damit Reduktion von Maschinenstillständen und damit eine effektivere Nutzung des Anlageninvest.
- Reduktion von Prozesskosten durch Wegfall manueller Bearbeitungszeit und Rüstaufwand
- Erhöhung der Flexibilität bei gleichbleibender Qualität
- Entlastung der Mitarbeitenden durch ergonomische Teilautomation und damit Verbesserung der Arbeitsplatzumfelds
- Stärkung des Teams “Mensch-Maschine” zur Steigerung der Wettbewerbsfähigkeit und Wirtschaftlichkeit
- Eröffnung neuer Märkte durch Automatisierbarkeit auch von hochindividuellen Produkten
- Stärkung der Resilienz durch höhere Prozessstabilität bei variierenden Anforderungen

Insbesondere für kleine und mittlere Unternehmen (KMU) entsteht durch die intuitive Bedienung und geringe Einstiegshürden ein realistischer Zugang zur industriellen Robotik – auch ohne eigene Automatisierungsexpertise. Der entwickelte und validierte generalisierte Lösungsansatz ist aufgrund seiner technologischen Kombination in vielen Anwendungsfelder der Fertigung anwendbar. Je nach Bedarf und Prozess kann der Funktionsumfang modular gestaltet werden – Von der einfachen Bedienerführung über Projektion bis hin zum interaktiven Programmieren von situativen Prozessteilschritten auf dem Bauteil selbst. Das erlaubt den Anwender auch diese Teil-Technologien in verschieden Ausprägungen an verschiedenen Fertigungsstellen einzusetzen.

Zusammenfassung

Das vorgestellte Assistenzsystem markiert einen signifikanten Fortschritt in der Mensch-Roboter-Kollaboration für die industrielle Fertigung. Durch die intelligente Kombination aus Sensortechnologie, KI-gestützter Bahnplanung und multimodaler Interaktion wird der wirtschaftliche Einsatz von Robotik auch bei Einzelanfertigungen und variantenreichen Prozessen möglich. Die Validierung in realen Anwendungsfällen zeigt: Eine intuitive, teilautomatisierte Programmierung direkt am Werkstück ist nicht nur technisch umsetzbar, sondern ein entscheidender Hebel für die Wettbewerbsfähigkeit zukünftiger Produktionssysteme.

Literatur

[1] Mitteramskogler, J. J., Widmoser, F., Ikeda, M., Pichler, A.; Homography-based industrial image stitching Insight Session, (ERF 2024)

[2] Ikeda, M., Widmoser, F., Stübl, G., Zafari, S., Sackl, A., & Pichler, A. (2023, October). An Interactive Workplace for Improving Human Robot Collaboration: Sketch Workpiece Interface for Fast Teaching (SWIFT). In Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing (pp. 191-194).

[3] Anish Pratheepkumar, Markus Ikeda, Andreas Pichler, Markus Vincze, Towards Robotic 3D Surface Processing with Global Local Neural Region Descriptor Fields, (ERF2024)

[4] Möhl, P., Pratheepkumar, A., Ikeda, M., Pichler, A.; Morphing Based Transfer of Demonstrated Surface Finishing (ISM24)

Open Source Cybersecurity Anomalieerkennung

Alexander Szönyi
SBA-Research gGmbH
aszoenyi@sba-research.org

Andreas Ekelhart
SBA-Research gGmbH
aekelhart@sba-research.org

Markus Klemen
SBA-Research gGmbH
mklemen@sba-research.org

Kurzzusammenfassung

Die digitale Souveränität gewinnt zunehmend an Bedeutung. In einem Use Case mit einem Unternehmen mit kritischer digitaler Infrastruktur wurde untersucht, inwieweit Open-Source-Lösungen bestehende Technologien im Bereich Security Operations ergänzen oder ersetzen können. Der Fokus lag dabei auf der frühzeitigen Identifizierung potenzieller Angriffe und Betriebsstörungen. Zu diesem Zweck wurden Open-Source-Werkzeuge wie Wazuh und OpenSearch in eine produktionsnahe Umgebung integriert und unter realen Rahmenbedingungen evaluiert. Neben den technischen Integrationsaspekten wurden auch regulatorische Anforderungen und relevante Standards berücksichtigt. Ein besonderes Augenmerk galt der Verarbeitung großer Datenmengen. Die Ergebnisse zeigen sowohl Potenziale zur Erweiterung bestehender Sicherheitsarchitekturen durch Open Source als auch Herausforderungen bei Skalierung, Integration und Betriebsstabilität. Daraus ergeben sich praxisnahe Empfehlungen für den Einsatz von Open-Source-basierten Lösungen im industriellen Umfeld sowie in anderen Organisationstypen.

Keywords: Security Operations, Anomaly Detection, Open-Source Security, Digital Sovereignty

Abstract

Digital sovereignty is becoming increasingly important. In a use case involving a company with critical digital infrastructure, we examined the extent to which open-source solutions can complement or replace existing technologies in the area of security operations. The focus was on the early identification of potential attacks and operational disruptions. To this end, open-source tools such as Wazuh and OpenSearch were integrated into a production-like environment and evaluated under real-world conditions. In addition to the technical integration aspects, regulatory requirements and relevant standards were also taken into account. Particular attention was paid to the processing of large amounts of data. The results show both the potential for expanding existing security architectures through open source and the challenges of scaling, integration and operational stability. This has resulted in practical recommendations for the use of open source-based solutions in industrial environments and other types of organisations.

Keywords: Security Operations, Anomaly Detection, Open-Source Security, Digital Sovereignty

Einleitung

Die Absicherung komplexer IT-Infrastrukturen gewinnt in digitalen Wertschöpfungsketten zunehmend an Bedeutung. Unternehmen mit kritischer digitaler Infrastruktur sehen sich wachsenden Anforderungen an die Überwachung und Analyse ihrer Datenströme gegenüber, um potenzielle Angriffe und Betriebsstörungen frühzeitig erkennen und abwehren zu können. Neben technologischen Herausforderungen müssen auch regulatorische Vorgaben, wie die EU-Richtlinien NIS1 und NIS2, sowie internationale Standards wie ISO/IEC 27001 berücksichtigt werden, um ein angemessenes Sicherheitsniveau sicherzustellen.

Vor diesem Hintergrund wurde in einem Use Case untersucht, inwieweit Open-Source-Lösungen bestehende Technologien im Bereich Security Operations ergänzen oder sogar ersetzen können. Der Schwerpunkt lag auf der Anomalieerkennung, einem zentralen Bestandteil moderner Sicherheitsarchitekturen, der es ermöglicht, Abweichungen von üblichen Betriebszuständen zu identifizieren und so potenzielle Angriffe frühzeitig zu erkennen. Um diese Fragestellung zu beantworten, wurden Open-Source-Komponenten wie Wazuh und OpenSearch in eine produktionsnahe Umgebung integriert und unter realen Rahmenbedingungen evaluiert.

Der betrachtete Anwendungsfall ist durch ein sehr hohes Datenaufkommen geprägt, das aus Cybersicherheits-Perspektive kontinuierlich überwacht werden muss. Die zu verarbeitenden Daten umfassen sowohl standardisierte Netzwerk- und Systemprotokolle als auch anwendungsspezifische Daten, die für das Kerngeschäft des Unternehmens von zentraler Bedeutung sind. Hierzu zählen Protokolle von Firewalls, Switches und Routern (über das Syslog-Protokoll), Netzwerkverkehrsdaten, Provider-spezifische Verbindungsdaten, Portalinformationen sowie Endpunkt-Telemetriedaten. Abhängig von der Netzwerkarchitektur entstehen dabei hohe Datenraten, die typischerweise in *Events per Second* (EPS) gemessen werden. Im vorliegenden Projekt wurde daher die Verarbeitung von Eingangsdatenströmen mit einer Rate von mehr als 100.000 bis 200.000 EPS untersucht, was einem Datenaufkommen von etwa 1 TB pro Tag entspricht (siehe Tabelle 1). Unter Berücksichtigung einer Aufbewahrungsfrist von 365 Tagen ergeben sich damit Speicher- und Verarbeitungsanforderungen im Bereich von mehreren Hundert Terabyte.

Tabelle 1. Geschätztes Datenaufkommen pro Tag.

Quelle	GB/Tag
Netflow Daten	100
Syslog Daten	150
Provider Daten	400
Endpoint Daten	100
Summe	**750**

Zur Umsetzung des Vorhabens wurden leistungsfähige Open-Source-Komponenten eingesetzt, die sich für den Aufbau und Betrieb moderner Security-Operations-Funktionalitäten eignen. Im Fokus standen insbesondere die Lösungen OpenSearch für die hochskalierbare Speicherung und Analyse von Logdaten sowie Wazuh für die Sammlung, Normalisierung und Korrelation sicherheitsrelevanter Ereignisse. Ziel des Projekts war es, die Leistungsfähigkeit dieser Technologien in einer produktionsnahen Umgebung zu evaluieren, Integrationshürden zu identifizieren und Handlungsempfehlungen für die Einführung von Open-Source-basierten Anomalieerkennungsplattformen abzuleiten.

Problemlösungsweg

Für die Evaluierung wurde ein methodischer Ansatz gewählt, der auf einer modularen Architektur und dem Einsatz virtualisierter Open-Source-Komponenten basiert. Ziel war es, unterschiedliche Datenquellen unter realen Bedingungen zu erfassen, vorzuverarbeiten und für Sicherheitsanalysen bereitzustellen. Dabei standen insbesondere die Skalierbarkeit der Datenerfassung, die Lastverteilung auf mehrere Instanzen sowie die zuverlässige Weiterleitung an Analyse-Cluster im Fokus. Die entwickelte Lösung basiert auf einer modularen Architektur, die verschiedene Open-Source-Komponenten integriert und eine flexible Skalierung ermöglicht (siehe Abbildung 1).

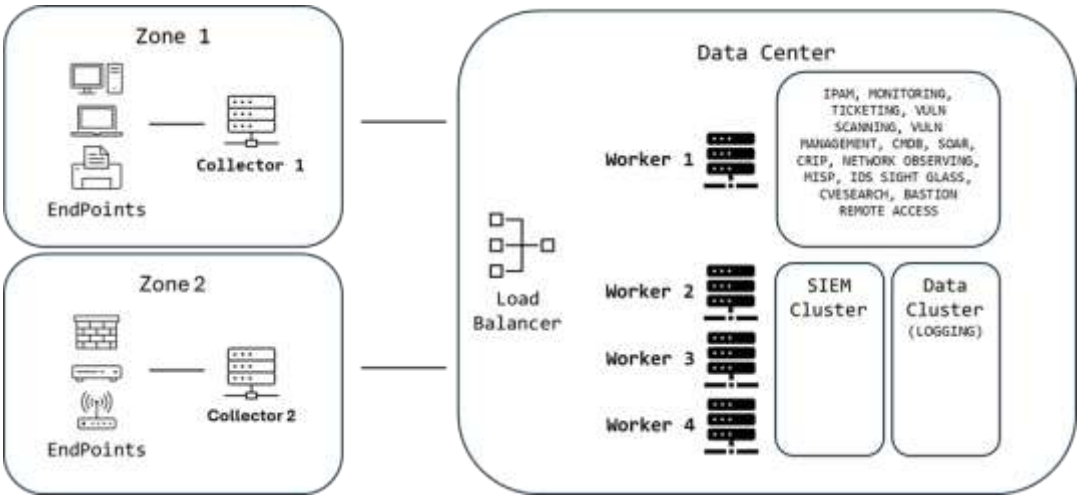

Abbildung 1. Beispielhafte Darstellung der Open Source-basierenden SOC-Lösung.

Zur Umsetzung der SOC-Funktionalitäten wurden mehrere Open-Source-Module integriert, die unterschiedliche Aufgaben wie Asset Management, Schwachstellenanalyse oder Netzwerküberwachung abdecken (siehe Abbildung 2).

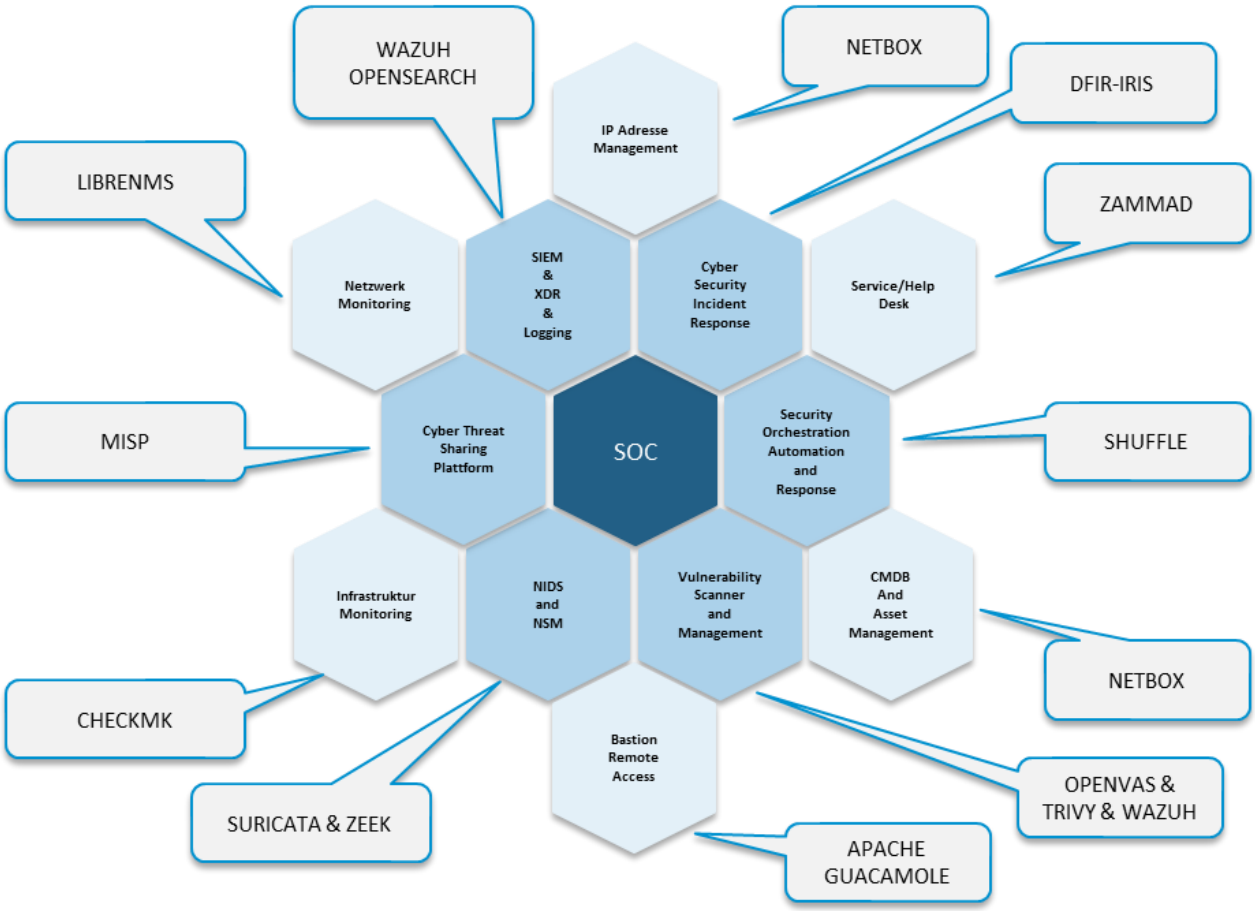

Abbildung 2. Darstellung zum Einsatz gekommen Module der Open Source basierenden SOC-Lösung.

Im Folgenden wird auf den Lösungsansatz zur Verarbeitung der unterschiedlichen Datenquellen eingegangen.

Syslog Daten

Für die Aufnahme von Syslog-Daten wurde ein Wazuh Agent als Eingangs-Collector gewählt, da Wazuh bereits eine integrierte Verarbeitungskette für diese Datenquellen bereitstellt. Der Agent empfängt die Datenströme und leitet diese an einen Wazuh Server Cluster weiter, der die Ereignisse analysiert und für nachgelagerte Sicherheitsprozesse aufbereitet (siehe Abbildung 3).

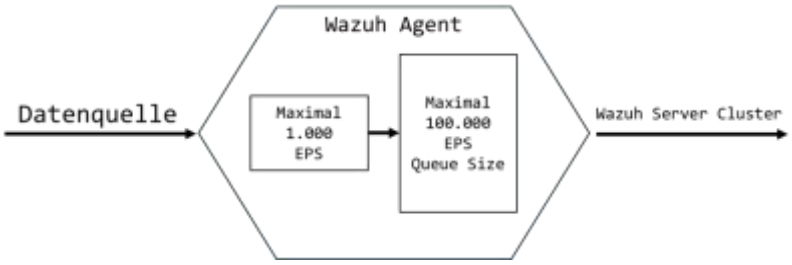

Abbildung 3. Die Darstellung zeigt den Datenfluss und die Limitierungen.

Ein einzelner Wazuh Agent unterliegt jedoch Leistungsgrenzen bezüglich der Anzahl verarbeitbarer Events per Second (EPS) sowie der maximalen Zwischenspeicherung von Ereignissen (Queue Size). Um eine Datenrate von mehr als 100.000 EPS verarbeiten zu können, wurden daher mehrere Wazuh Agents parallel innerhalb einer virtualisierten Container-Umgebung betrieben. Zur Lastverteilung kam ein Open-Source-basierter Load Balancer zum Einsatz, der den Datenstrom über dedizierte Netzwerkports auf die einzelnen Agenten verteilt (siehe Abbildung 4).

Im Testbetrieb konnten bis zu 20 parallel laufende Wazuh Agents auf einem einzelnen Host-System betrieben werden, ohne dass es zu Datenverlusten kam.

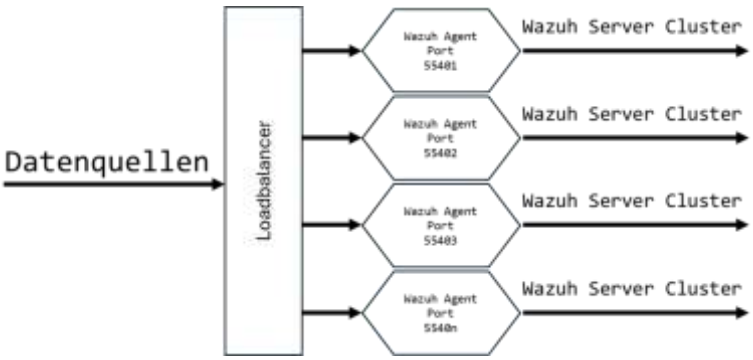

Abbildung 4. Die Darstellung zeigt den Datenfluss und die Aufteilung des Traffics auf mehrere Agenten.

Netflow Daten

Für die Verarbeitung von Netzwerkverkehrsdaten, insbesondere Netflow (IPv4/IPv6), wurde ein analoger Ansatz gewählt: virtualisierte Logstash-Instanzen dienten als Empfänger der Netflow-Datenströme und übergaben die aufbereiteten Daten an den OpenSearch-Cluster.

Provider Daten

Auch Provider-spezifische Verbindungs- und Betriebsdaten wurden mit Hilfe von virtualisierten Logstash-Instanzen gesammelt, normalisiert und zur weiteren Analyse in die zentrale OpenSearch-Datenhaltung integriert.

Ergebnis

Die Evaluierung hat gezeigt, dass Open-Source-Lösungen in der Lage sind, sehr große Datenmengen zuverlässig und ohne Datenverlust zu verarbeiten. Ein zentraler Eigenbeitrag des Projekts bestand darin, verschiedene Open-Source-Komponenten, darunter Wazuh, OpenSearch und virtualisierte Logstash-Instanzen, in einem modularen Framework zu kombinieren und auf die spezifischen Anforderungen des Anwendungsfalls anzupassen. Durch die flexible Architektur konnten unterschiedliche Datenquellen integriert, verarbeitet und in einer skalierbaren Testumgebung unter realistischen Bedingungen evaluiert werden. Dabei wurde nachgewiesen, dass sowohl standardisierte als auch geschäftsrelevante individuelle Daten mit vergleichsweise geringem Integrationsaufwand eingebunden werden können.

Ein wesentliches Merkmal des entwickelten Lösungsstacks ist die vollständige On-Premis-Verarbeitung aller sicherheitsrelevanten Informationen, sodass keine Daten das unternehmenseigene Rechenzentrum verlassen. Dadurch wird die digitale Souveränität der Informationen gewährleistet – ein entscheidender Faktor insbesondere für Betreiber kritischer Infrastrukturen. Die erzielten Ergebnisse stehen im Einklang mit den Zielen des aktuellen österreichischen Regierungsprogramms, das die Stärkung von Innovationskraft und Digitalisierung betont und die Vermeidung einseitiger technologischer Abhängigkeiten fordert. Aufbauend auf den gewonnenen Erkenntnissen bietet das entwickelte Framework Potenzial für weitere Anwendungsfelder, etwa die Einbindung zusätzlicher Datenquellen oder die Entwicklung automatisierter Reaktionsmechanismen. Darüber hinaus eröffnet die modulare Architektur Möglichkeiten für Kooperationen mit anderen Organisationen, um gemeinsame Sicherheitsstandards und offene Referenzarchitekturen zu etablieren.

Zusammenfassung

Das Projekt hat gezeigt, dass durch die Integration mehrerer Open-Source-Komponenten ein leistungsfähiges Framework zur Anomalieerkennung aufgebaut werden kann, dass auch bei sehr hohen Datenraten zuverlässig arbeitet und individuelle, geschäftsrelevante Datenquellen flexibel einbindet. Ein wesentlicher Vorteil liegt in der vollständigen On-Premis-Verarbeitung aller sicherheitsrelevanten Informationen, wodurch die digitale Souveränität gewährleistet wird. Darüber hinaus weist der Ansatz weitere Stärken auf: Er ist für Organisationen jeder Größe und Branche skalierbar, unterstützt die Einhaltung regulatorischer Vorgaben (z. B. NIS1/2, DORA) sowie internationaler Standards (z. B. ISO/IEC 27001, ISO 62443) und ermöglicht durch seine offene Architektur eine schnelle Implementierung. Der Einsatz von Open-Source-Lösungen stärkt zudem die heimische Wirtschaft und profitiert von einer globalen Community, die kontinuierlich Weiterentwicklungen vorantreibt. Damit sind die erzielten Ergebnisse für eine Vielzahl von Organisationstypen und Branchen von hoher Relevanz.

Für die Zukunft ergeben sich mehrere Forschungsschwerpunkte. Dazu gehören die Normalisierung von ISP-Log-Daten, um eine standardisierte und konsistente Datenbasis zu schaffen, sowie die Erweiterung der Anomalieerkennung um Machine-Learning-Methoden und ein spezifisches Endpoint-Profiling.

Literatur

[1] Wazuh, *Agent Client Buffer Einstellungen*. [Online]. Available: https://documentation.wazuh.com/current/user-manual/reference/ossec-conf/client-buffer.html. [Accessed: 30-Jul-2025].

[2] OpenSearch, *Logstash Plugin*. [Online]. Available: https://docs.opensearch.org/docs/latest/tools/logstash/index/.[Accessed: 30-Jul-2025].

[3] Heise Online, *Strafgerichtshof: Microsofts E-Mail-Sperre als Weckruf für digitale Souveränität*. [Online]. Available: https://www.heise.de/news/Strafgerichtshof-Microsofts-E-Mail-Sperre-als-Weckruf-fuer-digitale-Souveraenitaet-10387368.html. [Accessed: 30-Jul-2025].

[4] United States Department of Justice, *Division V—Cloud Act*. [Online]. Available: https://www.justice.gov/criminal/media/999391/dl?inline. [Accessed: 30-Jul-2025].

[5] Golem.de, Microsoft kann US-Zugriff auf EU-Cloud nicht verhindern. [Online]. Available: https://www.golem.de/news/anton-carniaux-microsoft-kann-us-zugriff-auf-eu-cloud-nicht-verhindern-2507-198283.html. [Accessed: 30-Jul-2025].

RAG is Not All You Need!

Raphael Pisoni
Software Competence Center Hagenberg
Raphael.pisoni@scch.at

Mario Pichler
Software Competence Center Hagenberg
Mario.pichler@scch.at

Kurzzusammenfassung

Retrieval Augmented Generation (RAG) hat sich zu einem Standardparadigma entwickelt, das es Large Language Models (LLMs) ermöglicht, Fragen unter Verwendung privater oder Echtzeit-Informationen zu beantworten. Durch das Abrufen relevanter Dokumente und deren Bereitstellung als Kontext für den Generator zielen RAG-Systeme darauf ab, faktenbasiertere und spezifischere Antworten zu erzeugen. In diesem Beitrag wird jedoch argumentiert, dass der Standard-RAG-Ansatz für komplexe Informationsbedürfnisse unzureichend ist. Es werden zwei grundlegende Schwächen identifiziert: Erstens erfasst der Retrieval-Schritt oft nicht den vollständigen Kontext, der für eine genaue Beantwortung einer Frage erforderlich ist, was dazu führt, dass unvollständige oder irreführende Informationen an den Generator weitergegeben werden. Zweitens kann selbst bei erfolgreichem Retrieval das schiere Volumen des relevanten Textes die Kapazität des *Context Windows* des Generators überschreiten, was eine Kürzung der Informationen erzwingt und das Modell daran hindert, eine umfassende Antwort zu synthetisieren. Um diese Mängel zu beheben, wird eine neuartige Methode vorgeschlagen: Retrieval Augmented Context Engineering (RACE). Diese Methode führt zunächst ein breites Retrieval potenziell relevanter Textausschnitte durch. Anschließend fungiert ein generatives Modell als intelligenter Filter, der entweder irrelevante Dokumente durch die Erzeugung eines einzigen „None“-Tokens schnell verwirft oder prägnante, auf die Frage ausgerichtete Zusammenfassungen der relevanten Dokumente erstellt. Diese kompakten Zusammenfassungen, die alle für die Frage relevanten Informationen enthalten, können dann gemeinsam in das *Context Window* des finalen Generators eingespeist werden. Dadurch wird sichergestellt, dass der Generator Zugriff auf die gesamte Bandbreite der notwendigen Informationen hat, ohne dass es zu einem Überlaufen des Kontext Windows kommt, was qualitativ hochwertigere und vollständigere Antworten erzeugt.

Keywords: Large Language Models, Retrieval Augmented Generation, Industrielle Anwendungen

Abstract

Retrieval Augmented Generation (RAG) has become a standard paradigm for enabling Large Language Models (LLMs) to answer questions using private or real-time information. By retrieving relevant documents and providing them as context to the generator, RAG systems aim to produce more factual and specific answers. However, this paper argues that the standard RAG approach is insufficient for complex information needs. We identify two fundamental weaknesses: first, the retrieval step often fails to capture the complete context required to answer a question accurately, leading to incomplete or misleading information being passed to the generator. Second, even when retrieval is successful, the sheer volume of relevant text can exceed the generator's context window capacity, forcing a truncation of information and preventing the model from synthesizing a comprehensive answer. To address these shortcomings, we propose a novel method, *Retrieval Augmented Context Engineering (RACE)*. This method first performs a broad retrieval of potentially relevant text snippets. Then, a generative model acts as an intelligent filter, either rapidly discarding irrelevant documents by generating a single "None" token or producing concise, question-focused summaries of the relevant ones. These compact summaries, which retain all question-relevant information, can then be collectively fed into the final generator's context window. This ensures that the generator has access to the full breadth of necessary information without context overflow, which is designed to produce higher quality and more complete answers.

Keywords: Large Language Models, Retrieval Augmented Generation, Industrial Applications

Introduction

The advent of Large Language Models (LLMs) has revolutionized natural language processing, but their inherent knowledge is static, limited to the data they were trained on. Retrieval Augmented Generation (RAG) [1] was introduced as an elegant solution to this problem, allowing LLMs to access and utilize external knowledge sources at inference time. The typical RAG pipeline is straightforward: given a user query, a retriever (often based on vector similarity search) [2] fetches a set of documents deemed relevant from a corpus. These documents are then concatenated with the original query and fed as context to a generator LLM, which produces the final answer.

This approach has proven effective for many tasks, particularly for fact-based questions where the answer is explicitly contained within a few retrieved passages. However, as we move towards more complex, multi-faceted questions, the limitations of this simple pipeline become increasingly apparent. The quality of the final generation is critically dependent on the quality of the retrieval, a process that is far from perfect. Furthermore, the physical constraint of the LLM's context window imposes a hard limit on the amount of information the model can "see" at once.

This paper contends that for complex queries, RAG is not enough. We will first delve into the dual problems of incomplete context retrieval and context window overflow. In its place, we propose Retrieval Augmented Context Engineering (RACE), an architecture that enhances the RAG framework by introducing a generative filtering and summarization step. This intermediate process ensures that the final generation stage receives a context that is both complete and concise, containing all the necessary information to formulate a high-quality answer without being constrained by context length.

The Shortcomings of Standard RAG

The efficacy of a standard RAG system hinges on two core assumptions: (1) the retriever can successfully find all the relevant information, and (2) the generator can fit all this relevant information into its context window. In practice, both of these assumptions frequently break down.

The Incompleteness of Retrieval

Modern retrieval systems, while powerful, are imperfect. Vector-based search, the most common method, excels at finding semantic similarity but can struggle with nuanced queries. A query might require synthesizing information from multiple, disparate sources that do not explicitly mention each other or the exact keywords in the query. For instance, a question like, "What was the market's reaction to Company X's sustainability report, considering the new regulations announced last quarter?" requires retrieving at least two distinct types of documents: financial news about Company X and documents detailing the new regulations.

A standard RAG system might retrieve one or the other, but rarely both, especially if the documents do not directly cross-reference each other. The retriever, optimized for finding a few "top-k" relevant chunks, often misses pieces of the puzzle. This leads to a generator that is working with an incomplete picture, resulting in answers that are partially correct but lack the depth and synthesis the user is seeking [3, 4]. The model might correctly describe the market reaction but fail to connect it to the new regulations, thereby failing to answer the user's full question.

The Challenge of Context Window Size

Even if the retriever could perfectly identify every single relevant document, a second, more practical problem arises: the limited context window of the generator LLM. While context windows have been growing, they remain a finite and valuable resource. Complex questions can easily have dozens of

relevant source documents or passages. Trying to concatenate all of this raw text into the prompt will quickly exceed the context limit of even the most advanced models.

The standard solution is to simply take the top-k retrieved documents that fit within the window. This truncation is a crude and often destructive form of compression. The (k+1)th document, which might contain a crucial piece of contradictory or clarifying information, is simply discarded. The model is forced to generate an answer based on an arbitrarily truncated subset of the relevant information. This can lead to answers that are not only incomplete but potentially incorrect, as the model lacks the full context to resolve ambiguities or identify nuances.

RACE: Retrieval Augmented Context Engineering

To overcome these limitations, we propose a two-step approach that augments the initial retrieval with a generative processing stage. This architecture is designed to ensure that the final context provided to the generator is both comprehensive and compact.

Step 1: Broad Initial Retrieval

The process begins similarly to standard RAG, but with a key difference: the initial retrieval is intentionally broad. Instead of aiming for a small, highly relevant "top-k" set, we retrieve a much larger number of potentially relevant text snippets. The goal of this stage is recall over precision; we want to cast a wide net to ensure that all possible pieces of relevant information are captured, even if it means including a significant number of irrelevant documents.

Step 2: Generative Filtering and Summarization

This is the core of our proposed solution. The large set of retrieved documents is not passed directly to the final generator. Instead, each document is individually processed by an intermediate generative model (which can be the same as, or a smaller version of, the final generator). For each document, this model is tasked with one of two actions, conditioned on the original user query:

1. **Discard**: If the document is irrelevant to the user's question, the model is trained to generate a single, predefined token, such as "None". This is an extremely fast and efficient operation, allowing the system to quickly discard a large volume of irrelevant text in a single forward pass and with minimal computational overhead.
2. **Summarize**: If the document is relevant, the model does not pass the full text along. Instead, it generates a concise summary of the document. Crucially, this is not a generic summary; it is a **question-focused summary**. The model is prompted to extract and synthesize only the information within the document that is directly pertinent to answering the user's query. This process distills the document down to its essential, relevant components, removing extraneous information.

For example, if the query is about the market reaction to a report, and a retrieved document is a 50-page sustainability report, the summarizer would not summarize the entire report. It would extract only the key metrics and statements that the market would likely react to, resulting in a summary that is perhaps only a few sentences long but contains all the necessary information from that source.

Step 3: Final Generation with Synthesized Context

After the filtering and summarization step, we are left with a collection of short, dense, question-focused summaries. Because these summaries are significantly shorter than the original documents, they can now be concatenated, along with the original query, and fit comfortably within the context window of the final generator LLM.

The final generator is now equipped with a context that has been specifically curated for the task at hand. It contains all the relevant information, drawn from a wide range of initial sources, and it is presented

in a compact, digestible format. This allows the model to perform the high-level synthesis and reasoning required to construct a comprehensive and nuanced answer, connecting the dots between the various pieces of summarized evidence.

Industrial Use Cases of RACE

The practical necessity for the RACE architecture is evident in complex challenges faced by industrial small and medium-sized enterprises (SMEs). Through our work on multiple TBI projects in the framework of the European Digital Innovation Hub (EDIH) AI5production, we have identified several recurring scenarios in both discrete manufacturing and process industries where standard RAG systems fall short. These use cases demonstrate a clear need for a more robust approach to information synthesis.

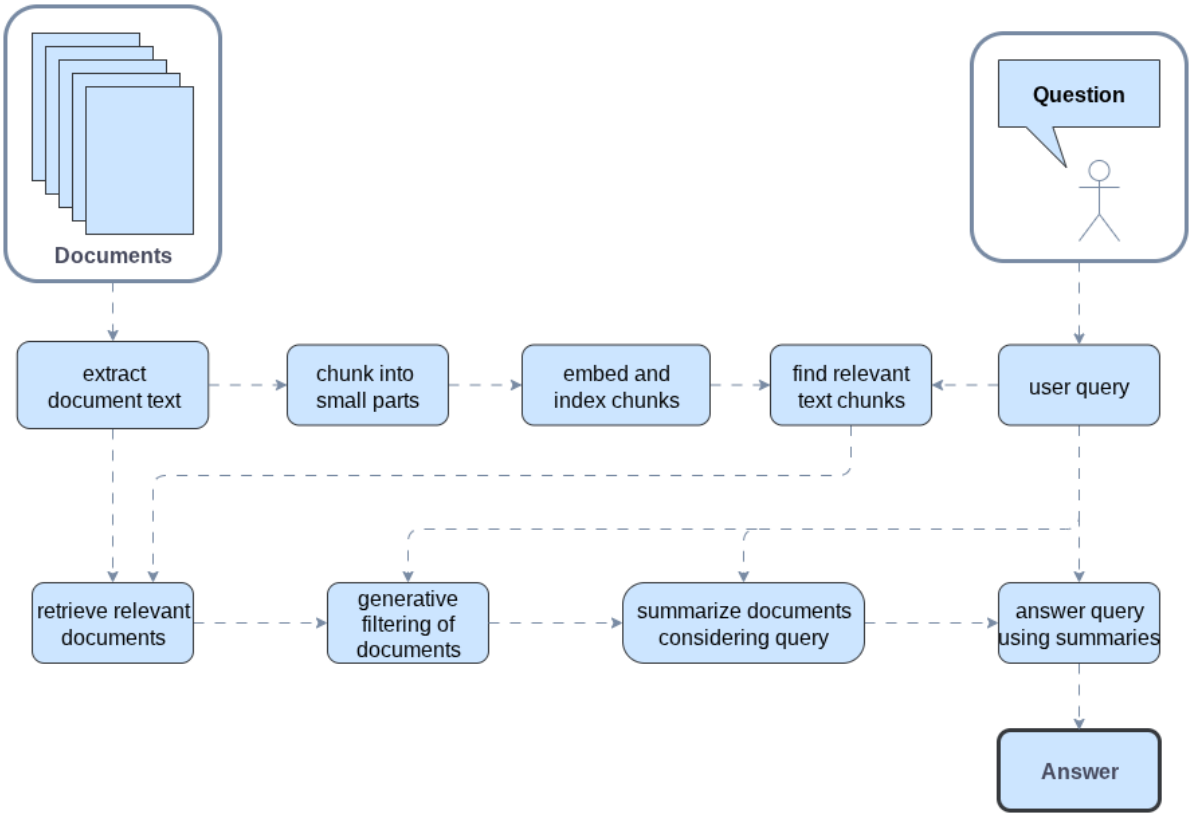

Figure 1. Retrieval Augmented Context Engineering Approach

These scenarios highlight a common industrial pattern: the need to make critical decisions by synthesizing incomplete, multi-source information. This is precisely where standard RAG fails and where the RACE framework provides a robust, practical solution.

Key industrial applications for RACE include:

- Complex Product Formulation: Developing a customer-specific chemical recipe, for instance, requires an expert to synthesize information from multiple technical manuals, safety data sheets, regulatory documents, and real-time material costs or availability. A standard RAG query might retrieve a base formula but miss a critical contraindication from a safety sheet or fail to suggest a cheaper, available alternative ingredient. RACE excels here by collecting all potentially relevant documents and providing concise, query-focused summaries, ensuring the final recommendation is safe, compliant, and cost-effective.
- Dynamic Process Optimization: To guarantee optimal plant configuration, an operator must consider specific material properties alongside equipment tolerances and process control parameters. A simple retrieval might not connect a new batch of raw material to the necessary

adjustments in machine settings, leading to suboptimal output or equipment wear. RACE can synthesize data from material datasheets and operational manuals to suggest holistic adjustments, ensuring peak performance.

- Resilient Supply Chain Management: When a required component is out of stock, finding a viable substitute involves searching internal inventories, supplier catalogues, and external technical databases. Standard RAG struggles to bridge these disparate sources to find and validate an alternative. RACE's broad retrieval and summarization can identify potential substitutes and synthesize the necessary information, including required modifications to tooling or configurations, to solve the problem.
- Automated Data Integration: For manufacturers dealing with small, custom batch sizes, the manual extraction of data from technical sheets and order forms into production plans is slow and error prone. RACE can automate this workflow by extracting and summarizing key parameters from diverse source documents, feeding a complete and structured context into production planning systems.

Conclusion and Future Work

The standard RAG pipeline, while a significant step forward, is a brittle solution for complex information needs. It suffers from both imperfect retrieval and the hard constraint of context window size. The Retrieval Augmented Context Engineering (RACE) approach presented in this paper directly addresses these two fundamental weaknesses. By separating the initial broad retrieval from the final generation with an intelligent generative processing step, we can construct a near-optimal context for the generator. This process ensures that all relevant information is identified and preserved, while irrelevant noise is efficiently discarded.

The result is a system that achieves significantly higher quality in both its retrieval (by retaining all relevant information) and its generation (by providing a complete, synthesized context). This method allows the LLM to move beyond simple fact extraction and engage in the kind of complex synthesis and analysis that users increasingly demand.

Future work will focus on optimizing the generative summarization step, exploring the trade-offs between different model sizes for this intermediate task, and developing methods to create structured summaries (e.g., tables, JSON, code), rather than purely textual ones, to further enhance the final generator's reasoning capabilities.

Further on and in relation with the work presented in this paper, we will follow, investigate and contribute to the development of manufacturing foundation models and their industries'-specific finetuning, also in combination with new developments of the Model Context Protocol [5] that allows to build two-way (real-time) connections between AI systems and required data sources. Another interesting recent development related to the introduced work in this paper is called Agentic Retrieval-Augmented Generation (Agentic RAG) that enhances traditional RAG by incorporating autonomous AI agents that use planning, reflection, and collaboration to manage complex, multistep tasks and dynamically refine retrieval strategies [6]. By utilizing those recent developments in our own future work, we will actively be paving the way for further advancement of Generative AI solutions in industrial settings, with a specific focus on their utilization opportunities by industrial SMEs.

With the presented RACE approach and the envisaged further investigations and developments, we feel confident to be prepared for similar TbI requests in the second EDIH phase as well as longer-term R&D collaborations with our SME partners.

References

[1] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Advances in Neural Information Processing Systems 33 (NeurIPS 2020), 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[2] V. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering," arXiv preprint arXiv:2004.04906, 2020. [Online]. Available: https://arxiv.org/abs/2004.04906

[3] N. F. Liu et al., "Lost in the Middle: How Language Models Use Long Contexts," arXiv preprint arXiv:2307.03172, 2023. [Online]. Available: https://arxiv.org/abs/2307.03172

[4] Z. Yue et al., “Inference Scaling for Long-Context Retrieval Augmented Generation”, Thirteenth International Conference on Learning Representations (ICLR 2025). [Online]. Available: https://openreview.net/pdf?id=FSjIrOm1vz

[5] Anthropic, "Model Context Protocol," *Anthropic News*, Nov. 04, 2024. [Online]. Available: https://www.anthropic.com/news/model-context-protocol

[6] A. Singh et al., “Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG”, arXiv preprint arXiv:2501.09136, 2025. [Online]. Available: https://arxiv.org/abs/2501.09136

Rekonstruktionspipeline für Low-Cost Aufnahmen zur Erstellung von Digital Twins

Attila Szabo
VRVis GmbH
szabo@vrvis.at

Lisa Kellner
VRVis GmbH
kellner@vrvis.at

Martin Riegelnegg
VRVis GmbH
riegelnegg@vrvis.at

Georg Haaser
VRVis GmbH
haaser@vrvis.at

Christoph Traxler
VRVis GmbH
traxler@vrvis.at

Kurzzusammenfassung

Neuartige Technologien in den Bereichen LiDAR und Photogrammetrie halten zunehmend Einzug in die Baustellenüberwachung und versprechen eine schnellere sowie effizientere Vermessung. Ihr volles Potenzial bleibt aber ungenutzt - fehlende Softwareintegration und unausgereifte Verarbeitungswerkzeuge bremsen den praktischen Einsatz. In diesem Beitrag stellen wir eine modulare Pipeline zur Rekonstruktion digitaler Zwillinge vor, die im Rahmen von drei EDIH-Projekten entwickelt wurde. Ziel ist die effiziente Verarbeitung von Baustellendatensätzen, insbesondere solcher, die mittels Low-Cost Vermessungstechnologien erfasst wurden. Die Pipeline kombiniert moderne Verfahren aus der Computer Vision mit künstlicher Intelligenz zur Bereinigung von Punktwolken und zur automatisierten Geometrierekonstruktion.

Keywords: Digital Twin, Punktwolke, Computer Vision

Abstract

New developments in LiDAR and photogrammetry are rapidly finding their way into construction site monitoring, promising faster and more efficient surveying processes. Yet, the broader potential of these technologies is still constrained by limited software integration and the lack of mature processing tools. This contribution introduces a modular pipeline for creating digital twins, developed in the context of three EDIH projects. The aim is to streamline the handling of construction site data, with a particular focus on datasets collected using affordable surveying equipment. The approach combines modern computer vision methods with AI-driven point cloud cleaning and automated geometry reconstruction, providing a practical step toward more accessible and scalable digital twin generation.

Keywords: Digital Twin, Point Cloud, Computer Vision

Einleitung

Auf Baustellen und in Infrastrukturprojekten werden zunehmend neuartige Technologien für die Low-Cost Erfassung von Messdaten eingesetzt. Insbesondere sind Endbenutzergeräte wie Smartphones gebräuchlich, die mit zusätzlichen Metainformationen wie GPS-Verortung und Zeitstempel als Dokumentationswerkzeug dienen. Außerdem versprechen Hersteller mancher Laserscanner wie XGRIDS [1] niedrigere Kosten und Zeitersparnis bei der Aufnahme.

Diese Geräte produzieren eine Vielzahl an Daten mit unterschiedlichen Qualitätsmerkmalen. So werden etwa moderne Smartphones mit einer LiDAR-Komponente ausgeliefert, die aber nur eine kurze Reichweite hat. Oder es beinhalten Punktwolken von günstigeren Laserscannern mehr Signalrauschen

als bei teureren Modellen. Herkömmliche Punktwolken-Prozessierungs-Pipelines sind für solche Fälle oft schlecht gewappnet. Fehlende integrierte Tools und Prozesse erzwingen händische Bereinigung und mühsame Auswertung der Datensätze.

Wir stellen eine Acquisition- und Analyse-Pipeline in mehreren Stages des Digital Twin Prozesses vor. Diese im Rahmen des EDIH-Programms durchgeführten Projekte reichern drei sukzessive Schritte in der Digital Twin Pipeline mit AI und Computer Vision Algorithmen an.

Problembeschreibung

3D Baustelle - Effizientes Scanning von Künetten mit iPhone-LiDAR

Dieser Beitrag exploriert das Potenzial zur Baustellendokumentation des handelsüblichen iPhones, und den darin verbauten Tiefensensoren.

In diesem Beitrag beinhaltet die Dokumentation einer Künette (ein dünner Graben). Das iPhone 15 Pro Modell, das für die Vermessung verwendet wird, beinhaltet einen LiDAR (Light Detection and Ranging) Sensor mit vergleichsweise niedriger Reichweite und Auflösung. Aufnahme- und Post-Processing-Methoden, um trotzdem bei der Vermessung der Künette auf akzeptable Genauigkeit zu gelangen, werden exploriert. Die iPhone-Rohdaten werden gegen Referenzscans, die zeitgleich mit einem professionellen TLS aufgenommen wurden, verglichen.

AI-gestützte automatische Bereinigung von Punktwolken

Auf Baustellen ist kontinuierliche Dokumentation bzw. Fortschrittsüberwachung mittels LiDAR oder Photogrammetrie wünschenswert. Allerdings führen neue Low-Cost und High-Speed Technologien oftmals zu Punktwolken, die erheblich mit ungewünschten Fehlerpunkten (“Noise Points”) kontaminiert sind. Um diesem Problem entgegenzuwirken müssen Noise Points aus der Punktwolke entfernt werden. Traditionell geschieht das in Punktwolkensoftwares, in denen wiederholt händisch Lasso-Selektionen aus mehreren Blickwinkeln aufgezogen werden um Noise Points in 3D präzise anzuwählen. Das ist mit großem Aufwand verbunden und kostet viel Zeit.

In diesem Beitrag wurden zwei komplementäre Techniken entwickelt, um den Zeitaufwand beim Bereinigen von Noise Points zu reduzieren. Eine einfache Artificial Intelligence (AI)-unterstützte Punktselektion kommt anstatt der Lasso-Selektion zum Einsatz. Weiters wird der Bereinigungsprozess mittels Punktstabilitätsanalyse aus mehreren Scans teilautomatisiert.

RebarTwin: AI-beschleunigte geometrische Segmentierung von Bewehrungsstrukturen

Bewehrungsstrukturen sind fundamentale Bestandteile von Betonkonstruktionen. Dabei ist wichtig, eine Qualitätsüberprüfung durchzuführen, um strukturelle Integrität der Betonkonstruktion zu gewährleisten. Die Qualitätskontrolle wird im herkömmlichen Falle manuell durchgeführt. Die Beauftragten müssen dazu große Anzahlen an dünnen Metallstangen zählen und Abmessungen mit Konstruktionsplänen vergleichen. Dieser Prozess ist zeitaufwändig und fehlergefährdet.

In diesem Beitrag wird der Aufwand für die Qualitätskontrolle von Bewehrungseisen durch automatisierte Softwareunterstützung verringert. Das Resultat ist eine End-To-End Pipeline für die Segmentierung und die Rekonstruktion von zylindrischen Strukturen in Punktwolken, welche automatisiert gemessen und gezählt werden können.

Methodik

3D Baustelle - Effizientes Scanning von Künetten mit iPhone-LiDAR

Die Vermessung fand bei Künetten an zwei Standorten statt, eine in Bogenneusiedl und eine in Groß-Enzersdorf. Es ergaben sich dreimal Scandaten:

- Hochpräziser Laserscan mit einem professionellen TLS.
- iPhone-LiDAR Rohdatenscan und Prozessierung in RTAB-Map [2].
- Photogrammetrie unter Verwendung der iPhone-Kamera.

Eine gesondert angefertigte iOS-App liefert alle zugänglichen Sensordaten aus der iPhone Sensor API als Datenstream. Diese Daten sind: Interne/Externe Kameraparameter, Farbfotos, Tiefenbild aus dem LiDAR Sensor, Segmentierungsmaske, stabile Features, und Kamera-Trajektoriedaten (GPS, Kompass).

Praxisorientierte Scan-Methodiken werden evaluiert, um Best Practices für das Scannen mit einem iPhone zu identifizieren. Dies erweitert den aktuellen State-of-the-Art, in dem hauptsächlich spezialisierte Strukturen rekonstruiert werden [5], selten aber die generelle Performance dieses Ansatzes beurteilt wird.

AI-gestützte automatische Bereinigung von Punktwolken

Eine der beiden Bereinigungsmethoden bietet vereinfachte Interaktivität, die andere läuft vollautomatisch. Die Techniken wirken synergistisch, da jeder Automatismus letztlich händisch überprüft und nachgebessert werden können muss.

Die traditionelle Lasso-Selektion, die mit vielen Mausklicks aus mehreren Betrachtungswinkeln erfolgt, wird durch eine semantische Selektion von Objekten mit nur einem Mausklick ersetzt. Dazu kommt eine gängige AI Bildsegmentierungsbibliothek zum Einsatz. Die 3D Punkte werden von einer Scanner-Position aus auf ein 2D Tiefenbild projiziert. Der Interessensbereich wird um den Mausklick-Punkt zentriert. Der Öffnungswinkel ist so kalibriert, dass das gewünschte Objekt etwa die Hälfte des Bildes einnimmt, was in diesem Kontext das beste Resultat geliefert hat. Diese Technik wurde aufgrund der hohen Performance und freien Verfügbarkeit gewählt, da die im State-of-the-Art eingesetzten 3D-Punktsegmentierungsansätze [6] aktuell schwer einzusetzen und auszuliefern sind.

Weiters wird die Puntkwolkenbereinigung durch die Analyse von Punktkonsistenz automatisiert. Die meisten Noise Points entstehen durch Bewegung oder Deformierung von Objekten. Solche Inkonsistenzen zwischen Scans werden durch Vergleich von Einzelbeobachtungen detektiert. Echte Oberflächen müssen dabei in mehreren Scans konsistent erscheinen, was Noise Points nicht tun. Für jeden Punkt ergibt sich ein Konsistenzmaß nach drei Faktoren. Punkte, die diese Kriterien verletzen, sind Noise Points und werden entfernt.

- Die Position im Raum in jedem Scan
- Die Konsistenz der rekonstruierten Oberflächennormalen in jedem Scan
- Die Anzahl der Scanner, die mit der Beobachtung übereinstimmen

RebarTwin: AI-beschleunigte geometrische Segmentierung von Bewehrungsstrukturen

Die Technik besteht aus zwei Komponenten: (1) Automatische Isolation von Bewehrungsstrukturen aus einer Punktwolke mittels Bildsegmentierungs-AI, und (2) Rekonstruktion von zylindrischen Primitiven. Daten-Input ist eine photogrammetrische Punktwolke, die bereits vorgesäubert wurde, etwa mit den vorher beschriebenen Ansätzen.

Als ersten Schritt werden Bewehrungsstrukturen automatisch aus der größeren Punktwolke segmentiert. Ein AI-Bildsegmentierungsnetzwerk erkennt Bewehrungseisen in Fotos. Für das Training wird eine Pipeline für synthetische Datenerzeugung eingesetzt, wodurch die kostspielige manuelle Datenannotation nicht notwendig ist, und feinere Kontrolle über den Trainingsprozess möglich wird.

Nach der Segmentierung der Bewehrungsstrukturen in der 3D Punktwolke werden die einzelnen Metallstangen als Zylinderprimitive rekonstruiert. Ein Region-Growing Algorithmus verfolgt kohärente zylindrische Strukturen durch die Punktwolke hindurch. Ausgehend von einem Startpunkt wird abwechselnd Extrapolation und Optimierung angewandt, um die genaue Form der Struktur anzunähern.

In diesem Anwendungsfall ist es außerdem wichtig, dass explizit die Biegungen in den Eisenstangen als Torus Segmente modelliert werden. Dieser Ansatz entspricht etwa dem State-of-the-Art [7], hebt sich aber durch die Einbettung in ein interaktives System von diesem ab.

Implementierung und Resultate

3D Baustelle - Effizientes Scanning von Künetten mit iPhone-LiDAR

Auf dem iPhone wird über ARKit eine ARSession gestartet, durch welche fortlaufend ARFrames aufgenommen werden. Pro Frame können Foto, Depthmap, interne und externe Kameraparameter abgegriffen werden. Mit der Open3D Library werden aus diesen Daten eine eingefärbte Punktwolke pro Frame erzeugt. Durch Verwendung der Position und Orientierung der Kamera können die einzelnen Frame-Punktwolken zueinander ausgerichtet werden. So entsteht eine Punktwolke der gesamten Künette. Die folgende Abbildung 1 zeigt das Foto und die dazugehörige Depthmap eines Frames.

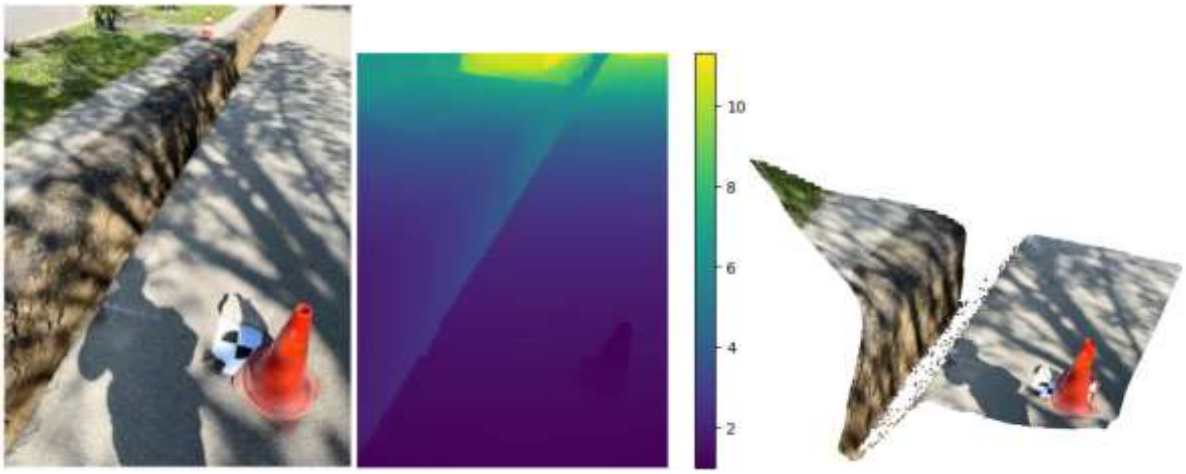

Abbildung 1. Mit dem iPhone aufgenommenes Foto (links), Tiefenbild (mitte), Punktwolke (rechts).

RTAB-Map ist eine Open Source Lösung für SLAM, die Loop Closing anhand von bildbasierten Features und Markern unterstützt. Als Proof-of-Concept wurden mehrere Aruco Marker ausgelegt und die Szene mit der RTAB-Map iOS App aufgenommen.

Mehrere praktische Aufnahmemethodiken wurden evaluiert. Da bei einer der Künetten eine Seite schwer zugänglich war, kamen sowohl die einseitige, als auch die beidseitige Begehung zum Einsatz. Die Evaluierung ist in folgender Tabelle 1 ersichtlich.

Tabelle 1. Vergleich verschiedener Aufnahmemethodiken mit dem iPhone.

Aufnahmeart	Resultat
Hochformat oder Querformat	Querformat fängt mehr Umgebung seitlich der Künette ein.
Nur auf einer Seite gehen	Eine Seite der Künette ist dünner aufgenommen.
Hingehen auf einer Seite, Retour auf der anderen	Beide Seiten der Künette dicht, aber Registeriungs-Drift.
Selfie-Stick	Liefert das beste Ergebnis, da man das iPhone in die Künette halten kann und das Abgehen auf einer Seite ausreicht.

In der Folgenden Abbildung 2 ist ein Querschnitt verschiedener Scans zu sehen die jeweils auf die Refernzpunktwolke (blau) registriert sind. Die Scans sind das iPhone LiDAR registriert ("1.+2. Hälfte"), unregistriert ("Original"), die kommerzielle Scan App "RTAB-Map" und die photogrammetrische Rekonstruktion von "COLMAP".

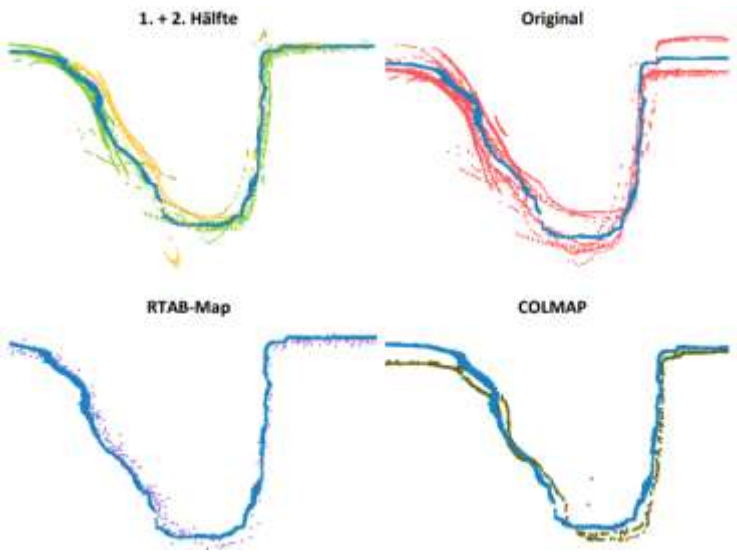

Abbildung 2. Vergleich von Aufnahmen via iPhone LiDAR registriert (links oben), unregistriert (rechts oben), Scan-App „RTAB-Map" (links unten) und Photogrammetrie (rechts unten), jeweils zur Referenzpunktwolke (blau).

Zusammenfassend kann gesagt werden, dass das iPhone Pro (ab Generation 12) mit seinen Sensoren für die 3D Baustelle im Fall der Künette geeignet ist. Die Tiefendaten des iPhones und die daraus rekonstruierten Punktwolken zeigten in visuellen wie auch in numerischen Vergleichen, dass sie eine Abweichung von nur einigen Zentimetern zum Referenzscan haben, was für den Fall der Künette ausreichen sollte.

AI-gestützte automatische Bereinigung von Punktwolken

Die frei verfügbaren Segmentierungsbibliothek SegmentAnything [3] kommt bei der AI-Punktwolkenbereinigung zum Einsatz. Mittels ONNX [4] wurden die AI-Gewichte aus dem Torch+CUDA Framework in eine herstellerunabhängige Umgebung exportiert. Dadurch kann die Segmentierung auch auf Kundenhardware ohne NVidia Grafikkarte ausgeführt werden. Ein Beispiel für einen so ausgewählten Menschen kann in der folgenden Abbildung 3 gesehen werden.

Abbildung 3. Mittels AI-Verfahren ausgewählter Mensch in einer Punktwolke.

Zur automatische Punktwolkenbereinigung wird eine Punkt in allen Panoramas verortet, die Normalen rekonstruiert, und ein entsprechendes Entscheidungskriterium angewandt. In der folgenden Abbildung

4 ist das Qualitätsmaß für jeden Punkt als Heatmap visualisiert. Man kann sehen, dass die automatische Selektion (oben) in wesentlichen Aspekten mit der manuellen Bereinigung (unten) übereinstimmt.

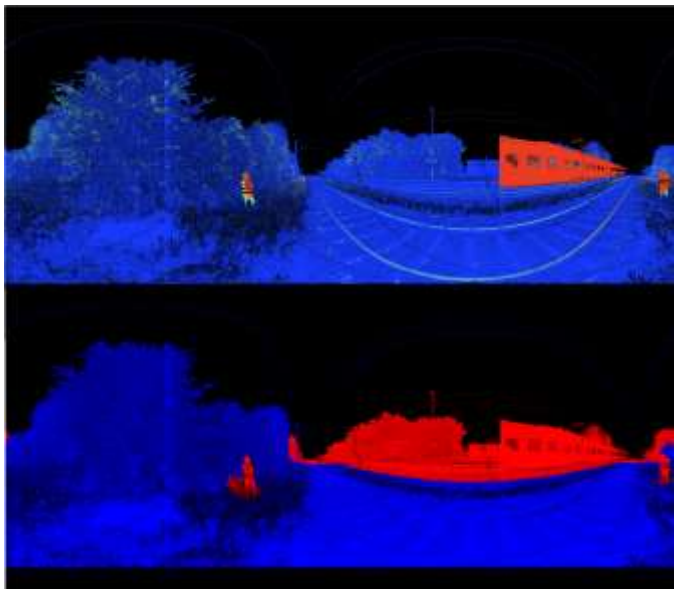

Abbildung 4. Oben: Durch automatische Punktwolkenanalyse ausgewählte Menschen und Zug im Scan. Unten: Händische Punktwolkenbereinigung.

Zusammenfassend ist zu beobachten, dass die vorgestellten Punktwolkenbereinigungstechniken den traditionellen Ansatz erheblich beschleunigen können. Teilweise kann die Bereinigung sogar genauer gemacht werden, etwa wenn beim händischen Lasso-Zeichnen zu viel angewählt wurde.

RebarTwin: AI-beschleunigte geometrische Segmentierung von Bewehrungsstrukturen

Deep Learning Netzwerke benötigen eine Vielzahl an annotierten Daten für das Training. Dafür dienen in Blender nachgebaute und zufällig in die Szene gesetzte Bewehrungseisen, die außerdem mit Hintergrundbildern versehen werden. Beispiele sind in der folgenden Abbildung 5.

Abbildung 5. Synthetisch generierte Fotos von Bewehrungseisen zum Training der AI-Segmentierung.

Als Netzwerkarchitektur wurde ein Feature Pyramid Networt (FPN) mit einem Mixed Transformer (MiT) als Backbone gewählt. Mit FPN werden Feature Maps in verschiedenen Größen mit nur wenig Mehraufwand erzeugt, wodurch das Netzwerk Objekte in unterschiedlichen Skalierungen erkennen und infolge dessen als Bewehrung, oder als keine Bewehrung, einordnen kann. Das Netzwerk konnte mit echten Testfotos eine gute Segmentierungsperformance erreichen, etwa wie in der folgenden Abbildung 6 erkennbar.

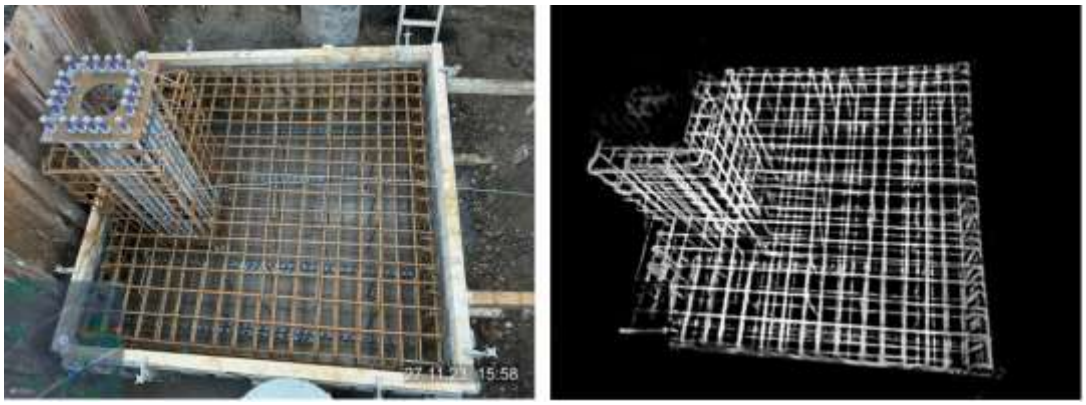

Abbildung 6. Links: Foto von Bewehrungskorb. Rechts: AI-Segmentierte Maske.

Die Rekonstruktion der Zylinder funktioniert in dem frühen Prototypstadium schon gut. Im Beispieldatensatz war der Rekonstruktionsfehler einer Stange bei etwa 0.8mm bei einem Durchmesser von 12mm. Dabei war die Skalierung der Punktwolke nicht ganz exakt (mit SLAM skaliert). Die verschiedenen Stangenmaße konnten eindeutig unterschieden werden. Eine Beispielrekonstruktion kann in der nächsten Abbildung 7 gesehen werden.

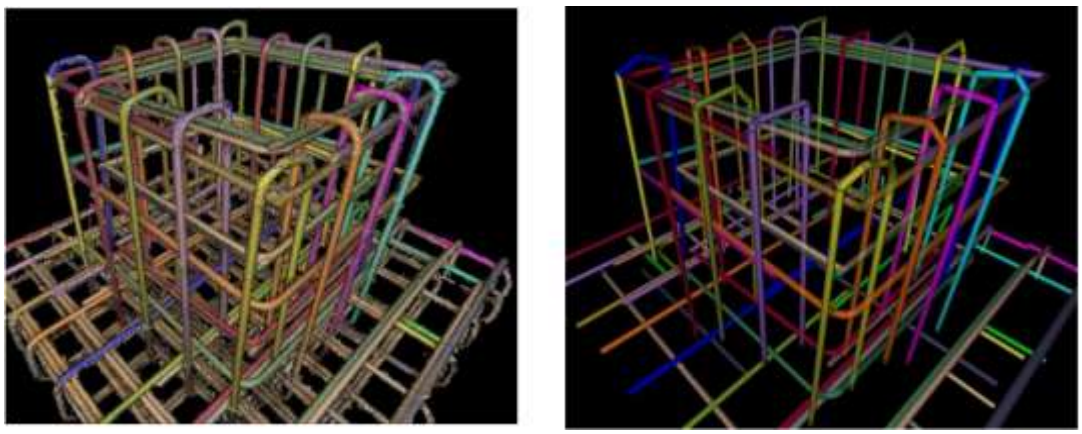

Abbildung 7. Mittels Region-Growing Verfahren rekonstruierte Bewehrungsstangen aus der Punktwolke.

Zusammenfassung

In diesem Beitrag wurde eine Aufnahme- und Rekonstruktions-Pipeline mit Low-Cost Geräten präsentiert. Da die Einzelprojekte im Rahmen des EDIH-Programms als Proof-of-Concept gedacht waren, zeigen sie bereits in diesem Stadium vielversprechende Möglichkeiten auf. Herkömmliche Konsumentengeräte eignen sich unter Einsatz entsprechender Software-Prozessierung als Aufnahmegeräte komplexer Daten mit höherer Qualitätsanforderung. Vor allem für Industrien die mit Digital Twins arbeiten und regelmäßige Bestandsaufnahme oder -kontrolle durchführen sollten die vorgestellten Techniken daher verwertbar sein.

Die in diesem Beitrag vorgestellten Ansätze sind in unterschiedlichem Umfang auf weitere Branchen übertragbar:

iPhone-LiDAR zur Datenerfassung: Der Einsatz handelsüblicher Smartphones ermöglicht hochwertige 3D-Aufnahmen durch nicht spezialisierte Fachkräfte und eignet sich daher etwa für Handwerksbetriebe wie Installations-, Maurer- oder Elektrogewerbe.

Bereinigung verrauschter Punktwolken: Daten aus unkontrollierten Umgebungen oder von kostengünstigen Sensoren sind oft durch Bewegungen oder Spiegelungen verfälscht. Die automatisierte Bereinigung bietet hier deutliche Vorteile, etwa im Mobile Mapping.

KI-gestützte Bildsegmentierung: Die Integration in Vermessungssoftware ist jetzt schon üblich und erlaubt die Auswertung von Bilddaten für vielfältige Anwendungen, darunter Baustellenüberwachung, industrielle Fertigung und medizinische Bildgebung.

Literatur

[1] XGRIDS. “XGRIDS-3D Reconstruction-Spatial Computing-Handheld Scanner.” Xgrids.com, 2025, www.xgrids.com/?page=geomatics. Accessed 29 July 2025.
[2] IntRoLab. “RTAB-Map.” RTAB-Map, introlab.github.io/rtabmap/. Accessed 29 July 2025.
[3] Kirillov, Alexander, et al. "Segment anything." Proceedings of the IEEE/CVF international conference on computer vision. 2023.
[4] Developers, Onnx Runtime. ONNX Runtime. 2021, https://onnxruntime.ai/.
[5] Mizutani, Tsukasa, and Shunsuke Iwai. "Improving construction site efficiency through automated progress monitoring of underground pipe installation sites using image color analysis of iPhone LiDAR camera data." *Developments in the Built Environment* 20 (2024): 100557.
[6] Zhou, Yuchen, et al. ‘Point-SAM: Promptable 3D Segmentation Model for Point Clouds’. *arXiv [Cs.CV]*, 2024, arxiv.org/abs/2406.17741. arXiv.
[7] Qiu, Rongqi, Qian-Yi Zhou, and Ulrich Neumann. "Pipe-run extraction and reconstruction from point clouds." *European Conference on Computer Vision*. Cham: Springer International Publishing, 2014.

Towards Optical Tracking for AR Assisted Oral Surgery

Philipp Fleck
VRVis GmbH
philipp.fleck@vrvis.at

Clemens Arth
VRVis GmbH
clemens.arth@vrvis.at

Kurzzusammenfassung

Jüngste Entwicklungen im Bereich Augmented Reality (AR) ermöglichen eine Vielzahl bisher ungeahnter Anwendungsgebiete und -bereiche. Insbesondere unterstützt AR-Technologie medizinische Aufgaben rund um den menschlichen Körper. In diesem Beitrag wenden wir AR-Technologie im zahnmedizinischen Bereich an, wo Zahnärzte oder Ärzte Mundhöhle und Zähne untersuchen. Wir beschreiben die ersten Schritte zur Posenschätzung einer Kamera anhand eines 3D-Modells von Teilen der Mundhöhle. Dabei nutzen wir Deep Learning und überlagern die Umrisse der oberen Gierung mit einem Live-Kamerabild. Wir identifizieren zwei Gruppen von Ansätzen: modellbasiertes, merkmalsloses Tracking und deren medizinische Anwendungen. Wir konzentrieren uns detaillierter auf den letztgenannten Ansatz und konzentrieren uns auf merkmalsbasiertes und merkmalsloses Tracking. Der erste Schritt basiert auf einem einfachen Kantenerkennungsalgorithmus für das Modell und das tatsächliche Livebild. Der zweite Schritt besteht darin, ein dreidimensionales (3D-)Bild der Zähne mit dem Livebild zu überlagern. Neben dem medizinischen Bereich ist unser Ansatz auch auf andere texturlose starre Objekte anwendbar, bei denen Konturbilder extrahiert werden können.

Keywords: model-based tracking, dental, feature-less tracking, augmented reality, oral surgery

Abstract

Recent developments in augmented reality (AR) allow for a wide range of never-thought-of applications and domains. In particular, AR technology allows for the support of medical tasks around the human body. In this work, we apply AR technology to the dental domain, where a dentist or physician investigates the oral cavity and teeth. We describe the first steps towards pose estimation of a camera with respect to a 3D model of parts of the oral cavities, leveraging deep learning and overlaying the outlines of the upper yaw onto a live camera feed. We identify two groups of approaches, namely, model-based, feature-less tracking and their medical applications. We focus on the latter type of approach in more detail, and we focus on feature-based tracking and featureless tracking. The first step is based on a simple edge detection algorithm on the model and the actual live image. The second step is to overlay a three-dimensional (3D) image of the teeth onto the live image feed. Besides the medical domain, our approach is applicable to other texture-less rigid objects where contour images are extractable.

Keywords: model-based tracking, dental, feature-less tracking, augmented reality, oral surgery

Introduction

Recent developments in augmented reality (AR) allow for a wide range of never-thought-of applications and domains. In this work, we focus on a human-centered use case to support medical tasks around the human body. In particular, we apply AR technology to the dental domain, where a dentist or physician investigates the oral cavity and teeth. This is of high interest to the industry and scientific community, as many "off-the-book tasks" require extensive preparation. Such tasks include the preparation of implants, which sometimes involve the construction of a custom-made drilling jig for each patient. Furthermore, the result of such procedures is hardly presentable to patients since there is no "on-patient" visualization using AR or similar technologies.

Figure 1. Conceptual drawing of the proposed approach. On the left, a 3D model of the upper jaw is shown. As a result, on the right the teeth outlines are overlaid on an image during live mode.

The ability to spatially address an individual tooth or other feature in the oral cavity enables a new range of applications. Ranging from pre-flight training, over in-situ visualization, to navigational support during operation, we can only guess the added value. However, we can still draw conclusions from other similar fields with manual hand operation. With the rise of Industry 4.0 and the digitization trend, the manufacturing and maintenance of machinery created the need for more state-of-the-art guidance. In this regard, systems that support operators performing a task were found to be very beneficial. The use of AR visualizations in situ was thus manifested to empower workers in different domains.

Our aim is now to translate these findings into the dental domain and to identify a possible future roadmap towards oral surgery that uses AR support. We therefore describe the first steps towards pose estimation of a camera with respect to a 3D model of parts of the oral cavity, leveraging deep learning and overlaying the outlines of the upper yaw onto a live camera feed (see Figure 1 for an examplary picture).

Related Work

The first steps in AR already demonstrated the potential to precisely track and enhance the perceived world [1], [2]. However, due to technical limitations, markers [3], [4], [5] were the first choice to robustly fulfill the core requirement for AR, namely the 6-degree-of-freedom pose tracking, i.e. robustly estimating the geometric relationship between a camera, respectively, display, and the world at frame rate. Although marker tracking is still common due to its robustness especially in challenging industrial environments, it is nowadays considered old-fashioned, and placing markers is also impractical in certain use cases. Therefore, over the last 15 years research has gradually shifted towards markerless tracking. In that regard, we can identify two groups of approaches, namely a) feature-based tracking and b) feature-less tracking. Since teeth have few distinctive features, we focus on the latter type of approach in more detail, i.e. on model-based, feature-less approaches and their medical applications.

Model based tracking

In contrast to marker- or natural feature-based approaches, model-based methods try to estimate the pose based on the given shape, i.e. a geometrical model. Such techniques are especially interesting for textureless rigid objects. Rehg et. al. [6] demonstrated how to overcome self-occlusion with model-based tracking already back in 1995. Comport et. al. [7] introduced a robust model-based version for robot vision. Wüst et. al. [8] utilized lines in a multi-hypothesis procedure to estimate the pose of the objects. Later, with the support of CAD models as source model, Wüst et. al. [9] were able to track industrial models, but also estimate the initial pose [10] in a more robust way. Usually such approaches are based on lines, hollows, depth, or shape and try to minimize the distance between the source representation (i.e. the model) and the observation.

Iterative closest point algorithms [11], voting algorithms [12] or similar are used to establish correspondences through matching, which then serve as input to pose estimation. With the development of powerful handheld devices, real-time image processing tracking of CD models [13] on mobiles became possible. Seo et. al. [14] present a direct method based on monocular RGB images. More recently, MOLTR [15] was capable of localizing, tracking, and reconstructing multiple objects from live daily scenes. Model-based tracking also finds its application in the automotive or autonomous driving field, where sparse LIDAR data are used to track pedestrians [16].

Our presented approach takes inspiration from the approaches of the last 20 years and applies them to the dental setting, in which camera tracking for AR applications has not been widely evaluated or established.

Medical AR

Within the medical domain, assistive systems [17] became more and more important in recent years. Since medical personnel often undergoes varying stress levels [18] it is beneficial to reduce mental loads [19] for certain tasks such as navigation and to free up brain capacity for tasks such as problem solving. This problem is observed in all medical disciplines, but also in dental surgery [20].

Wang et. al. [21] presented an AR based system using a halfway mirror setup to create visual overlays for the operator on the stationary patient. An example was to highlight an individual tooth within the oral cavity. Other approaches, such as Karner et. al. [22] or Weber et. al.[23] propose complete head tracking based on CT-Scan or instant data, which are hardly applicable to dental use cases. One particular aspect in dentistry are deformations caused by instruments such as mouth openers or similar devices. The oral cavity is mainly composed of soft and flexible tissue, so the assumptions about rigidity that are the core of most tracking approaches hold only for bones and teeth at most.

Dental reconstruction [24] and facial reconstruction [25] are used for preparation and real-time navigation. Other methods [26] used reconstructions to create precise implants for structural defects. In the dental industry, intraoral scanners and jaw tracking systems are increasing in popularity [27].

However, our work strives to close the gap between well-known tracking algorithms in AR and special conditions in the dental domain, by creating new algorithms adopted to the human characteristic and the connected application cases.

Methodology

Traditional model-based tracking approaches consist of two major parts: a) the model creation and b) tracking. The acquisition step takes care of creating the trackable model. Depending on the features and technology used for the tracking, the model has to be prepared accordingly. For example, as Wüst et. al. [13] presented, the lines and other information are extracted from the CAD data to be used at runtime. Such algorithms require a known initialization pose of the model since there are only limited possibilities. Either the search space becomes large and extensive search has to be performed to initialize the tracking algorithm, or more commonly, the user aligns a predefined initial pose with the real world and reduces the search space by a magnitude.

Model Creation

Our assumption is the existence of an oral cavity scan, which is available as a 3D model (Figure 1, left) for further tracking purposes. Because such scans are very precise and up to scale, the model can directly serve as the basis for further data extraction and preparation stages for the subsequent tracking algorithm.

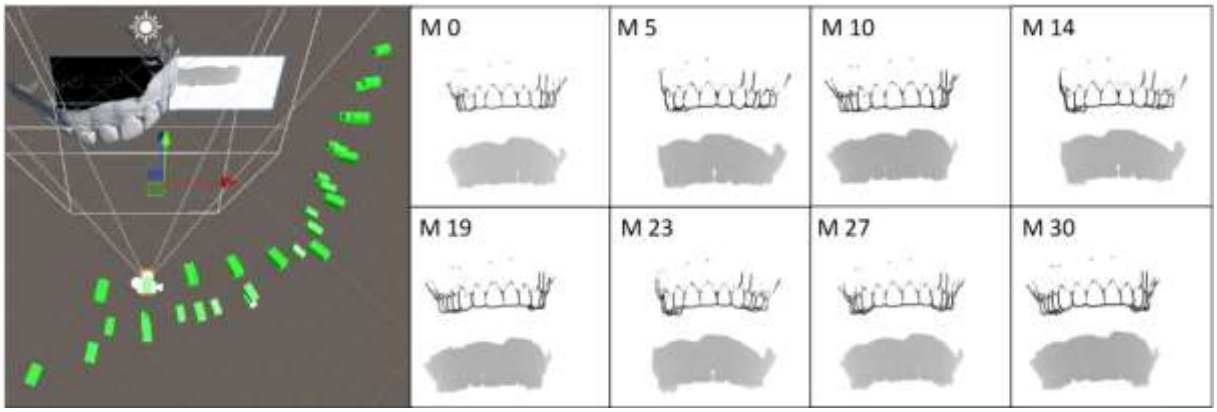

Figure 2. The 3D reconstruction of the upper jaw is loaded in Unity (left) and poses are sampled from different orientations (green). The resulting silhouette images and depth renderings are shown on the right "M 0" related to the first model pose.

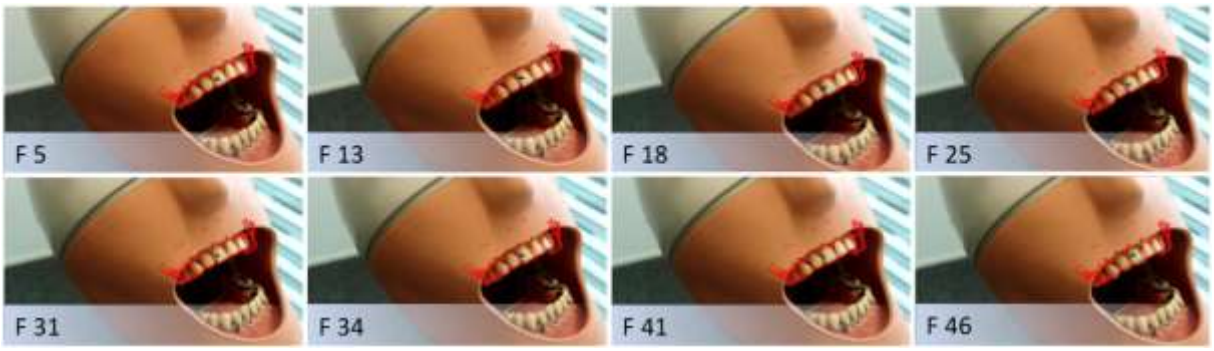

Figure 3. Tracking result over selection of frames of an 80 Frame long sequence. The prediction stays on target, and tries to optimize its alignment each frame.

We import the scan into a game engine, i.e. Unity, to create a variety of 2D images. The scan consists of the upper and lower jaw, the teeth, and the surrounding soft tissue. The images generated mimic the appearance of parts of the oral cavity as seen from a camera during tracking; therefore, we can extract any set of distinctive features that are known from long-time CV research; however, the major issue here is the absence of such features.

We therefore opt for an approach based on deep learning. We focus on the upper jaw for the moment and create a large number of artificial images which capture the appearance within a given hemispheric space. Although we do not use depth information, we also create a set of depth images alongside the 2D images of the model as seen from a known pose. As described in the work of Rad and Lepetit [28], we train a model to infer the pose based on a newly seen image. Some examples of the extracted images are shown in Figure 2. Note that training can be done in a very fast way because of the overall low number of images and the fact that we only need to estimate the pose of a single model (as opposed to an entire set of different objects).

Tracking

The tracking is implicitly solved by the network; i.e. the inference of the network is actually a pose of the object with respect to the camera position. The network is also able to infer intermediate poses, which means that the discrete space of samples and poses used for training becomes a continuous space for tracking; however, because the network is still trained from a defined number of samples, we add an optimization step for the pose, as we assume that the camera is handheld during tracking.

Figure 4. Selected frames out of a 50 frames sequence showing alternating error-alignment occurrences either due to drift or overshoot.

Figure 5. Visual comparison of the extracted input frame contour (green) against the predicted model (red). The limited light conditions within the mouth cavity cause a reduced amount of contours towards the back side of the jaw leading to inaccurate predictions in frame 15 and 20 (red).

We use a simplistic Kalman filter to model the camera motion. The pose representation, although based on only 6 degrees of freedom, is usually given as a 3x4 matrix, encapsulating a 3x3 matrix for the rotation and a 1x3 column vector for the translation. This representation cannot be easily used within an optimization approach. We therefore transform the pose into the *SE3* Lie space, which forms a differentiable manifold. The Kalman filter smoothes the motion path between previous and new inferences. For visualization, the resulting pose is transformed back into it's original format.

Visualization

To assess the overall fit, we use a simplistic edge detection algorithm on the model and the actual live image. This enables us to render the most prominent contours from our model and visually compare those with the contours in the live image. This is facilitated by importing the camera images into Unity and rendering the model contours into the images using a custom shader.

Experiments

We conducted several experiments using a dummy model of a patient's head. The 3D model of the upper jaw was created and processed as previously described. The overall development of our prototype is at an early stage; therefore, we only show a sequence of frames from a single test video. Also note that only the upper jaw was considered so far, and tracking the lower jaw (which may essentially be subject to a different type of motion compared to the rest of the patient's head) is considered future work.

In Figure 3 we show a selection of frames out of an 80-frame test sequence. Our prediction is capable of following the movement, but also re-adjusting just when the error increases. In such cases, the algorithm tries to minimize the distance to the observation by running the alignment step in a larger area. As with many prediction algorithms, such as Extended Kalman Filters, edge cases such as abrupt changes in direction can cause "overshooting", which is then compensated for the frame after occurrences as depicted in Figure 4.

In Figure 5, we compare the contour extracted from the input stream with the predicted model. Looking closely at frame 15 and 20, we identify a shorter contour towards the inside of the mouth cavity, due to lighting and blurriness. However, this causes the prediction to become inaccurate in the lower

distribution in the upper jaw. In a real-world scenario, a dentist would ensure good lighting conditions, which further helps reduce blur. In the future, we plan to gather more test data (also of human participants) for a more in-depth evaluation under real-world conditions.

Discussion and Conclusion

In this work, we described the first steps towards optical tracking of parts of the oral cavity based on deep learning. Besides the medical domain, other industrial applications, such as, part tracking, training, assembly guidance, are also meaningful domains. Similarly, VisionLib [13] focuses on CAD-based data in manufacturing for AR-based tracking applications. Furthermore, AR use-cases [29] demonstrated recently there benefit when enhancing the real-world with additional useful data.

We roughly outlined our initial approach and showed a couple of experimental results. Despite being a valid approach to the given problem, we found several deficiencies in our current prototypical implementation. First, the framework used for the initial evaluation is no longer considered state-of-the-art and was also originally built for a different use case. In deep learning research, it is important to strip the design and ultimate functionality of a neural network to its core requirements, both to save computational power and to increase accuracy at the same time. Therefore, we consider the application of more recent concepts [30] as the next step forward.

A second deficiency was found in the use of autofocus cameras, which are regularly used today without much reflection. However, with respect to the use case at hand, a system should ultimately provide both precise registration and accurate video overlay. This can hardly be achieved in setups where introspection or control of the camera used for live capture is not possible, \ie the varying focal length of the camera is known in the subsequent visualization procedure. We therefore consider the use of cameras with a fixed focus to be a better choice and contrary to what we used in our previous experiments.

Finally, the system is not applicable in real time. Future efforts concern the use of more lightweight hardware such that a practical implementation is possible without large-scale GPU power or even server resources.

Acknowledgment

This work was funded by the European Digital Innovation Hub (EDIH) under contract No. 101083472. However, views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.

References

[1] S. Henderson und S. Feiner, „Exploring the benefits of augmented reality documentation for maintenance and repair“, *IEEE transactions on visualization and computer graphics*, Bd. 17, Nr. 10, S. 1355–1368, 2010.

[2] S. Gauglitz, B. Nuernberger, M. Turk, und T. Höllerer, „World-stabilized annotations and virtual scene navigation for remote collaboration“, in *Proceedings of the 27th annual ACM symposium on User interface software and technology*, 2014, S. 449–459.

[3] H. Kato, „Inside ARToolKit“, in *1st IEEE international workshop on augmented reality toolkit*, 2007.

[4] D. Wagner und D. Schmalstieg, „Artoolkitplus for pose tracking on mobile devices“, 2007.

[5] M. Fiala, „ARTag, a fiducial marker system using digital techniques“, in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005, S. 590–596 Bd. 2. doi: 10.1109/CVPR.2005.74.

[6] J. M. Rehg und T. Kanade, „Model-based tracking of self-occluding articulated objects“, in *Proceedings of IEEE International Conference on Computer Vision*, 1995, S. 612–617. doi: 10.1109/ICCV.1995.466882.

[7] A. I. Comport, É. Marchand, und F. Chaumette, „Robust model-based tracking for robot vision“, in *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)*, IEEE, 2004, S. 692–697.

[8] Wuest, H. and Vial, F. and Stricker, D, „Adaptive line tracking with multiple hypotheses for augmented reality“, in *Fourth IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR '05)*, IEEE, 2005, S. 62–69.
[9] H. Wuest und D. Stricker, „Tracking of industrial objects by using cad models“, *Journal of Virtual Reality and Broadcasting*, Bd. 4, 2007.
[10] H. Wuest, F. Wientapper, und D. Stricker, „Adaptable model-based tracking using analysis-by-synthesis techniques“, in *Int. Conference on Computer Analysis of Images and Patterns*, Springer, 2007, S. 20–27.
[11] D. Chetverikov, D. Svirko, D. Stepanov, und P. Krsek, „The trimmed iterative closest point algorithm“, in *2002 International Conference on Pattern Recognition*, IEEE, 2002, S. 545–548.
[12] D. Chang und S. Hashimoto, „An inverse voting algorithm for Hough transform“, in *Proceedings of 1st International Conference on Image Processing*, IEEE, 1994, S. 223–227.
[13] H. Wuest, T. Engelke, F. Wientapper, F. Schmitt, und J. Keil, „From CAD to 3D Tracking — Enhancing & Scaling Model-Based Tracking for Industrial Appliances“, in *IEEE Int. Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*, 2016, S. 346–347. doi: 10.1109/ISMAR-Adjunct.2016.0114.
[14] B.-K. Seo und H. Wuest, „A direct method for robust model-based 3D object tracking from a monocular RGB image“, in *Computer Vision–ECCV 2016 Workshops, Proc., Part III 14*, Springer, 2016, S. 551–562.
[15] K. Li, H. Rezatofighi, und I. Reid, „MOLTR: Multiple Object Localization, Tracking and Reconstruction From Monocular RGB Videos“, *IEEE Robotics and Automation Letters*, Bd. 6, Nr. 2, S. 3341–3348, 2021, doi: 10.1109/LRA.2021.3061080.
[16] P. Held, D. Steinhauser, A. Koch, T. Brandmeier, und U. T. Schwarz, „A Novel Approach for Model-Based Pedestrian Tracking Using Automotive Radar“, *IEEE Transactions on Intelligent Transportation Systems*, Bd. 23, Nr. 7, S. 7082–7095, 2022, doi: 10.1109/TITS.2021.3066680.
[17] R. Nagyné Elek und T. Haidegger, „Non-technical skill assessment and mental load evaluation in robot-assisted minimally invasive surgery“, *Sensors*, Bd. 21, Nr. 8, S. 2666, 2021.
[18] D. Yu *u. a.*, „Quantifying intraoperative workloads across the surgical team roles: room for better balance?“, *World journal of surgery*, Bd. 40, S. 1565–1574, 2016.
[19] A. Plessas, R. Witton, M. Paisi, F. Ellwood, und E. McColl, „Using the Mental Health Wellness Framework in dental practice“, *Bdj Team*, Bd. 9, Nr. 10, S. 10–12, 2022.
[20] M. Maleki Roveshti, M. Raei, und F. Valipour, „Examining Mental Workload and Incidence of Musculoskeletal Abnormalities of Dentists During Surgery“, *Archives of Rehabilitation*, Bd. 25, Nr. 2, S. 232–247, 2024.
[21] J. Wang *u. a.*, „Augmented Reality Navigation With Automatic Marker-Free Image Registration Using 3-D Image Overlay for Dental Surgery“, *IEEE Transactions on Biomedical Engineering*, Bd. 61, Nr. 4, S. 1295–1304, 2014, doi: 10.1109/TBME.2014.2301191.
[22] F. Karner *u. a.*, „Single-Shot Deep Volumetric Regression for Mobile Medical Augmented Reality“, in *Multimodal Learning for Clinical Decision Support and Clinical Image-Based Procedures*, Cham: Springer International Publishing, Jan. 2020, S. 64–74.
[23] M. Weber, D. Wild, J. Kleesiek, J. Egger, und C. Gsaxner, „Deep Learning-Based Point Cloud Registration for Augmented Reality-Guided Surgery“, in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, 2024, S. 1–5. doi: 10.1109/ISBI56570.2024.10635333.
[24] I.-B. Pavaloiu, N. Goga, I. Marin, und A. Vasilateanu, „Automatic segmentation for 3D dental reconstruction“, in *2015 6th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, IEEE, 2015, S. 1–6.
[25] W. Zielonka, T. Bolkart, und J. Thies, „Towards metrical reconstruction of human faces“, in *European conference on computer vision*, Springer, 2022, S. 250–269.
[26] J. Li *u. a.*, „Back to the roots: Reconstructing large and complex cranial defects using an image-based statistical shape model“, *Journal of Medical Systems*, Bd. 48, Nr. 1, S. 55, 2024.
[27] M. Revilla-León, D. E. Kois, J. M. Zeitler, W. Att, und J. C. Kois, „An overview of the digital occlusion technologies: Intraoral scanners, jaw tracking systems, and computerized occlusal analysis devices“, *Journal of Esthetic and Restorative Dentistry*, Bd. 35, Nr. 5, S. 735–744, 2023.
[28] M. Rad und V. Lepetit, „BB8: A Scalable, Accurate, Robust to Partial Occlusion Method for Predicting the 3D Poses of Challenging Objects without Using Depth“, in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, IEEE Computer Society, 2017, S. 3848–3856. doi: 10.1109/ICCV.2017.413.
[29] A. Stanescu *u. a.*, „Error Management for Augmented Reality Assembly Instructions“, in *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2024, S. 690–699. doi: 10.1109/ISMAR62088.2024.00084.
[30] B. Wen, W. Yang, J. Kautz, und S. Birchfield, „FoundationPose: Unified 6D Pose Estimation and Tracking of Novel Objects“, in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA: IEEE Computer Society, Juni 2024, S. 17868–17879. doi: 10.1109/CVPR52733.2024.01692.

AI5 PRODUCTION

AGENDA

15. OKTOBER 2025
14:30 - 21:00
JKU LINZ, OIC STUFENFORUM
ALTENBERGER STR. 69
4040 LINZ

14:30 — **EINLASS & GET TOGETHER**
Austausch mit Expert:innen, Poster & Live Use Cases

15:00 — **BEGRÜSSUNG**
- Rektor Univ.-Prof. Dr. Stefan Koch (JKU Linz)
- Rektorat TU Wien
- Vertreter:in der Europäischen Kommission

15:15 — **KEYNOTE – STRATEGIEN FÜR UNTERNEHMEN IM ZEITALTER VON KI UND INNOVATION**

15:45 — **LÖSUNGSANSÄTZE IN DER PRODUKTION (I)**
- ***Innovationsimpulse zur Digitalisierung von KMUs***
  Barbara Krumay & Manuel Mühlburger, JKU Linz
- ***Künstliche Intelligenz in der Produktion***
  Franz Sümecz, TU Wien & Barbara Krumay, JKU Linz
- ***RAG is Not All You Need! – Echtzeitinformationen im Kontext mit KI verarbeiten***
  Raphael Pisoni & Mario Pichler, Software Competence Center Hagenberg GmbH
- ***Digitale mensch-zentrierte Assistenzsysteme.***
  Harald Bauer, Gerhard Ebenhofer, Christoph Heindl & Gernot Stübl, PROFACTOR GmbH

16:45 — **PAUSE**

17:00 — **LÖSUNGSANSÄTZE IN DER PRODUKTION (II)**
- ***Rekonstruktionspipeline für Low-Cost Aufnahmen zur Erstellung von Digital Twins***
  Georg Haaser, Lisa Kellner, Martin Riegelnegg, Attila Szabo & Christoph Traxler, VRVis
- ***Open Source Cybersecurity Anomaliedarkennung.***
  Andreas Ekelhart, Alexander Szönyi & Markus Klemen, SBA Research GmbH
- ***Intelligente Chatbots zur Abfrage von Unternehmenswissen – ALLEX***
  Lukas Pichlmann, LKR Leichtmetallkompetenzzentrum Ranshofen
- ***Benefits für produzierende Unternehmen durch das Europäische Innovationsökosystem***
  Roland Sommer & Stefanie Wendelits, Plattform Industrie 4.0

18:00 — **VON AI5PRODUCTION ZU AI5INNOVATION**
WEITERE DREI JAHRE ALS AI5INNOVATION BIS 2028
- ***Fortsetzung von AI5production als AI5innovation bis 2028***
  Franz Sümecz, TU Wien, Coordinator AI5innovation
- ***Bundesminister Dr. Wolfgang Hattmannsdorfer***
  Gemeinderat Almir Balihodzic *& weitere Vertreter:innen aus Politik und Wirtschaft zur Bedeutung der Künstlichen Intelligenz für Österreich in wirtschaftlich herausfordernden Zeiten, der Stärkung des Technologiestandortes Linz durch gezielte Innovationsinitiativen, der Rolle der EDIHs der European Digital Innovation Hubs für Unternehmen und das Land Oberösterreich und die Europäische Ausrichtung der Künstlichen Intelligenz*

18:30 — **NETWORKING MIT BUFFET & LIVE-MUSIK**

Co-funded by the European Union
EDIH European Digital Innovation Hubs Network
Bundesministerium Wirtschaft, Energie und Tourismus
FFG

AI 5 INNOVATION

Ab 11/2025

Co-funded by the European Union

European Digital Innovation Hubs Network

Bundesministerium Wirtschaft, Energie und Tourismus

FFG Forschung wirkt.